Linear mixed models in sensometrics

Kuznetsova, Alexandra

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Linear mixed models in sensometrics

Alexandra Kuznetsova

Kongens Lyngby 2015
PhD-2015-374

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3351
compute@compute.dtu.dk
www.compute.dtu.dk PhD-2015-374

Summary (English)

Today's companies and researchers gather large amounts of data of different kind. In consumer studies the objective is the collection of the data to better understand consumer acceptance of products. In such studies a number of persons (generally not trained) are selected in order to score products in terms of preferences. In sensory studies the aim is the collection of the data to better describe products and differences of the products according to a number of sensory attributes. Here trained persons, so-called assessors, score the products in terms of different characteristics such as smell, taste, texture, sound - depending on the aim of a study. It is a common approach in both studies to consider persons coming from a larger population, which, from the statistical perspective, leads to the use of mixed effects models, where consumers/assessors enter as random effects (Lawless and Heymann, 1997).

Mixed effects models have been used extensively in analysis of both consumer and sensory studies. However frequently too simplistic models are considered, important effects are not accounted for and as a consequence important information is not gained or analysis leads to improper conclusions. The focus of this project is to propose a methodology for analyzing more complex models together with tools facilitating the methodology. This was accomplished by contributing to the mixed effects ANOVA modelling in general and specifically applied to sensory and consumer studies through a series of papers and software tools facilitating the developed methodologies. The primary advantage of the ANOVA approach is that it gives confidence intervals and significance tests for the various effects including the background variables used in the model and consequently a fast and reliable assessment and ranking of the importance of different factors.

There exists today very little easily available methodology and software which supports consumer studies with both sensory properties and background information related to health benefits, environment and user-friendliness. In close collaboration with the industrial partners an open-source software tool ConsumerCheck was developed in this project and now is available for everyone. will represent a major step forward when concerns this important problem in modern consumer driven product development. Standard statistical software packages can be used for some of the purposes, but for the specific problems considered here and for the typical users in industry, these programs are far from satisfactory. Therefore, the ConsumerCheck software represents a novel source of information for all quality-oriented industries. The effect is improved procedures for product development and hence improved quality of decision making in Danish as well as international food companies and other companies using the same methods.

The two open-source R packages **lmerTest** and **SensMixed** implement and support the methodological developments in the research papers as well as the ANOVA modelling part of the ConsumerCheck software. The **SensMixed** package is a package for semi-automated analysis of sensory and consumer studies within linear mixed effects framework. The **lmerTest** package supports tests for linear mixed effects models fitted with the `lmer` function of the **lme4** package (Bates et al., 2013). While **SensMixed** is closely connected with sensometrics field, the **lmerTest** package has developed into a generic statistical package. Reference manuals accompany these R packages.

Summary (Danish)

Virksomheder og forskere samler i dag store mængder af forskellig slags data. I forbrugerundersøgelser er målet at indsamle data for en bedre forståelse af forbrugernes accept af og præferencer for produkterne.. I sådanne undersøgelser er en række personer (normalt ikketrænede) valgt for at score produkter med hensyn til præferencer og/eller accpetance. I sensoriske undersøgelser er målet at indsamle data til bedre at beskrive produkter og forskelle i produkter i forhold til en række sensoriske egenskaber. Her scorer trænede personer, såkaldte bedømmere, produkterne i form af forskellige karakteristika såsom lugt, smag, konsistens, lyd - alt efter formålet med undersøgelsen. Det er en fælles tilgang i begge typer studier at betragte personerneso m kommende fra en større population, som således fra et statistisk perspektiv fører til brugen af "mixed effect"-modeller, hvor forbrugerne/bedømmerne bliver medtaget som tilfældige effekter.

"Mixed effect" modeller har allerede i udstrakt grad været brugt i analysen af både forbrugerundersøgelser og sensoriske undersøgelser. Men ofte betragtes alt for simple modeller, hvor der ikke tages højde for vigtige effekter og som en konsekvens kan vigtige oplysninger være misset og/eller analysen ført til forkerte konklusioner. Fokus i dette projekt er at foreslå en metodik til at analysere mere komplekse modeller sammen med værktøjer, der understøtter metodikken. Dette er opnået ved at bidrage til "mixed effects" ANOVA modellering generelt og specifikt anvendt på sensoriske undersøgelser og forbrugerundersøgelser gennem en række artikler og softwareværktøjer, der understøtter de udviklede metodikker. Den primære fordel ved ANOVA tilgangen er, at den giver konfidensintervaller og signifikanstests for de forskellige effekter, herunder de supplerende baggrundsvariable, der anvendes i modellen og dermed en hurtig og pålidelig vurdering og ranking af vigtigheden af forskellige faktorer.

Der eksisterer i dag meget få lettilgængelige metodikker og software, som understøtter forbrugerundersøgelser med både sensoriske egenskaber og baggrundsvariable i relation til sundhedsmæssige fordele, miljø og brugervenlighed. I tæt samarbejde med de industrielle partnere er et open-source software værktøj "ConsumerCheck" blevet udviklet i dette projekt, og er nu tilgængelig for alle. Dette er et stort fremskridt, hvad angår denne vigtige udfordring i moderne forbrugerdrevet produktudvikling. Standard statistiske softwarepakker kan anvendes til nogle af de formål, men for de specifikke problemer der betragtes her og for de typiske brugere i industrien, er disse programmer langt fra tilfredsstillende. Derfor repræsenterer "ConsumerCheck" software en ny kilde til information for alle kvalitetsorienterede industrier. Effekten er forbedrede procedurer for produktudvikling og dermed forbedret kvalitet af beslutningsprocessen i såvel danske som internationale fødevarevirksomheder og andre virksomheder, der bruger samme fremgangsmåder.

De to open source R pakker **lmerTest** og **SensMixed** implementerer og understøtter de metodiske udviklinger i forskningsartiklerne samt ANOVA modelleringsdelen af ConsumerCheck software. **SensMixed** pakken er en pakke til semi-automatiseret analyse af sensoriske undersøgelser samt forbrugerundersøgelser indenfor rammerne af de "mixede lineære modeller". **lmerTest** pakken understøtter tests for lineære mixed effects modeller modelleret med `lmer` funktionen fra **lme4** pakken (Bates et al., 2013). Mens **SensMixed** er tæt forbundet med sensometriområdet har **lmerTest** pakken udviklet sig til en generisk statistisk pakke. Referencemanualer medfølger disse R pakker.

Preface

This thesis was prepared at the Section of Statistics and Data Analysis of department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU), in partial fulfilment of the requirements for acquiring the Ph.D. degree in Applied Mathematics.

The thesis deals with linear mixed effects models in sensometrics. Sensometrics is the scientific area that applies mathematical and statistical methods to problems from sensory and consumer science. The main focus is on developing statistical methods, models and software tools for linear mixed effects models applied in sensory and consumer science.

The thesis consists of six research papers, two R packages documented by their reference manuals and one stand-alone software. For the full list of works and publications associated with the Ph.D. see page vii

Lyngby, -2015

Alexandra Kuznetsova

List of papers, software and other contributions

This thesis is based on six scientific papers, two add-on packages for the statistical programming language R (R Development Core Team, 2015); two reference manuals support these software packages, one stand-alone software ConsumerCheck.

[A] **Kuznetsova, A.**, Christensen, R. H. B., Bavay, C., Brockhoff, P. B. (2014). *Automated Mixed ANOVA Modeling of Sensory and Consumer Data*. Food Quality and Preference 40: 31–38. doi:10.1016/j.foodqual.2014.08.004.

[B] **Kuznetsova, A.**, Brockhoff, P. B., Christensen, R. H. B. *lmerTest package: Tests in Linear Mixed Effects Models*. (submitted to Journal Of Statistical Software)

[C] **Kuznetsova A.**, Amorim I., Brockhoff P. B., Lima R. R. *Analysing sensory data in a mixed effects model framework using the R package SensMixed* (intended for Food Quality and Preference Journal)

[D] Tomic, O., **Kuznetsova, A.**, Brockhoff, P. B. , Graff, T., Naes, T. *ConsumerCheck: a software for analysis of sensory and consumer data* (submitted to Journal of Statistical Software)

[E] Brockhoff P. B., Amorim I., **Kuznetsova A.**, Søren Bech, Lima R. R., *d-prime interpretation of a standard linear mixed model results* (accepted in Food Quality and Preference Journal)

[F] Bavay, C., Brockhoff, P. B., **Kuznetsova, A.**, Maître, I., Mehinagic, E., Symoneaux, R. (2014). *Consideration of Sample Heterogeneity and in-Depth Analysis of Individual Differences in Sensory Analysis.*Food Quality and Preference 32: 126–31. doi:10.1016/j.foodqual.2013.06.003.

The following two R packages have been implemented:

- The **lmerTest** package http://cran.r-project.org/web/packages/lmerTest/index.html
- The **SensMixed** package https://r-forge.r-project.org/R/?group_id=1433

The following two reference manuals accompany these R packages:

[G] **Kuznetsova, A.**, Brockhoff, P. B., Christensen, R. H. B. *lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-20.*

[H] **Kuznetsova, A.**, Brockhoff, P. B., Christensen, R. H. B. *SensMixed: Mixed effects modelling for sensory and consumer data. R package version 2.0-6.*

The following tutorial accompany the **SensMixed** R package:

[I] **Kuznetsova, A.** *Tutorial for the R-package SensMixed.*

The following stand-alone tool was developed in collaboration with ConsumerCheck partners http://consumercheck.co/

In addition to a number of presentations and posters the following contributed talks have been given on material included in this thesis:

- Automated Mixed ANOVA Modelling of sensory and consumer data *The 11th Sensometrics Conference, 10th - 13th July 2012, Rennes, France*
- Different tests on lmer objects (of the lme4 package): introducing the lmerTest package *UseR!: The R User Conference 2013, July 10-12 2013. University of Castilla-La Mancha, Albacete, Spain*

- SensMixed: A New R-Package for user friendly mixed model ANOVA for multi attribute Sensory and Consumer data *The 12th Sensometrics Conference, 29th July to 1st August 2014, Chicago, USA*

The following papers has been published or prepared in collaboration with other researchers during the PhD period. These papers are not part of the methodological developments included in this project and will not be further addressed:

- Cècile Bavay, Ronan Symoneaux, Isabelle Maître, **Alexandra Kuznetsova**, Per Bruun Brockhoff, Emira Mehinagic, *Importance of fruit variability in the assessment of apple quality by sensory evaluation*, Postharvest Biology and Technology, Volume 77, March 2013, Pages 67-74, ISSN 0925-5214, http://dx.doi.org/10.1016/j.postharvbio.2012.11.005.
- Line H. Mielby, Heidi Kildegaard, Sidsel Jensen, **Alexandra Kuznetsova**, Nina Eggers, Barbara V. Andersen, Per B. Brockhoff and Derek V. Byrne *Sensory perception of fibre and flavour added fruit drinks sweetened with stevia and the effect on Consumers Liking, Wanting and Sensory Satisfaction* intended for food research international
- Line H. Mielby, Sidsel Jensen, Heidi Kildegaard, **Alexandra Kuznetsova**, Nina Eggers, Barbara V. Andersen, Per B. Brockhoff and Derek V. Byrne *Effect of means of context evocation and type of evoked context on consumers' response towards fruit drinks* intended for Food Quality and Preference

Acknowledgements

First and foremost I would like to thank my supervisor Per Bruun Brockhoff. Thank you for all your help, support, guidance that you have provided all these years. It was a great honour to be your Ph.D. student. I can only hope that, at least at some degree, I have fulfilled your expectations regarding my work here at DTU.

A special thank goes to Rune Haubo Christensen, who was a great co-supervisor. Thank you for your enthusiastic supervision and encouragement and for always being supportive with relevant feedback. I am very proud that I had a chance to work with you and to have you as a mentor.

Many thanks to the ConsumerCheck partners, especially to Tormod Naes and Oliver Tomic. You have been very helpful and our discussions on various topics related to sensometrics have been really rewarding. Thank you for warmly welcoming me at Nofima.

I would like to thank all my colleagues from the Statistics and Data analysis section for providing an inspiring working environment. Many thanks to PhD fellows for all the lunch and coffee breaks.

I would like to thank my parents for their love and help and for always being there for me.

Most of all I would like to thank my husband Roman for all the love and support during all these years and especially during the tough time of writing the thesis. Thanks to my two children, Andrey and Mariya, for all their smiles and kisses.

Contents

Chapter 1

Introduction

In most buying situations today the consumer takes a large number of product properties into account. First of all the sensory properties of the product itself must be liked, but the modern consumer is generally interested in several more aspects. Important additional product properties (added values) of special interest in modern society are benefits for the health, user-friendliness of the product and aspects related to the environment (related to raw materials, ingredients used, production process and packaging). While food science journals have published lots of studies investigating how sensory attributes affect consumer liking, consumer science has focused more on how different cognitive and emotional aspects influence product liking. Methodology for studying the effects of sensory properties on consumer liking is well established under the name of preference mapping McEwan (1996) and methodology for investigating other factors such as health labelling and user-friendliness is usually known under the name of conjoint analysis (Gustafsson et al., 2003). In order to understand how added values interact with the sensory properties of the product itself, it is crucial for consumer driven product development that both aspects are investigated simultaneously. Methodology for this type of combined studies is today in high demand internationally, but very little tailor made methodology exists. No user-friendly software system is available.

When both sensory attributes and context variables are involved in the design of the study, classical Analysis of Covariance (ANCOVA, see e.g. Weisberg

(1985)) methodology incorporating both qualitative and quantitative variables in a single regression equation can sometimes be useful. When combining sensory and other information, however, this type of methodology suffers from serious problems. The main reason is that a typical sensory profile consists of many attributes (typically between 15 and 20) which in most cases are also highly collinear. Combining such variables with a set of categorical information or context variables in situations with possibly strong interactions is a complex task: Standard ANCOVA methods will either not be possible to use or provide very unstable solutions. This is particularly true when the number of objects is moderate or small as it usually is in product development (6-8 products). The present project develops methodology which is a one step further meeting these challenges.

In the sensory studies trained persons, so-called assessors, score the products in terms of different characteristics such as smell, taste, texture, sound - depending on the aim of a study. The fact that generally people experience different perceptions from the same influences in their senses, use differently the scale in scoring the products leads to several levels of variation. Accounting to different sources/levels of variation constitutes a special challenge to a sensory scientist. Frequently, due to the lack of easy-to-use methodology together with the tool facilitating it, too simplistic methods are chosen and important systematic variation in the data is not captured, which sometimes leads to misleading conclusions. In this project one of the focuses was to provide a novel methodology to handle complex settings, where important information is captured and the results are easily interpreted.

In sensometrics field it has been a consensus to consider assessors as well as consumers as randomly selected from a larger population, which leads to the use of mixed effects models, where assessors/consumers are treated as random effects (Lawless and Heymann, 1997). Even though mixed effects models are more and more used for analysis of sensory data, still frequently too simplistic models are considered, so that not all important information is always captured nor accounted for. In sensory studies much work has been done on modeling differences between assessors and accounting for them through analysis of variance (ANOVA) and not that much on modelling different carry-over effects. Frequently they are simply neglected. Bavay et al. (2014) showed that it is essential to model also carry-over effects especially in studies, where products are prone to biological heterogeneity (e.g. fruits, vegetables, cheeses, etc.). In such studies variations in the data may be due to assessor differences and/or product heterogeneity. In Kuznetsova et al. (2015) the methodology was developed that facilitates model-building approach and helps to find a parsimonious mixed effects model that captures an important information from consumer as well as sensory data.

In sensory studies there always exist individual differences between assessors in their way of assessing the products. Some of the differences are related to the assessors sensitivity, others can be related, for example, to the use of the intensity scale. Although different use of scale is not directly related to the quality of the assessors, it is important to being able to detect it and account for it in the modelling analysis if possible (Naes et al., 2010). Brockhoff et al. (2015) proposed a so-called mixed assessor model (MAM), where the scaling effect is modelled. However, the model introduced there together with the tool facilitating the analysis of the model had some limitations. For instance, the model can not account for potential carry-over effects. And, as pointed out in Bavay et al. (2014), in a variety of studies, it is important to model also the carry-over effects. In this work an extended version of MAM is proposed together with the tool facilitating the analysis.

It is of particular importance to develop tools for simple and visual interpretation of the results based on the statistical methods developed. In particular this is important for ensuring efficient communication between the statistician and the user of the methods. In sensory studies assessors score the products for a wide range of attributes. In order to perform comparative analysis between the attributes, a so-called multi-attribute analysis is of demand. In Nofima Mat (2008) the comparative multi-attribute plots are constructed and presented in a compact user-friendly way. However there are limitations on the use of the mixed effects models there. In Kuznetsova et al. (2013b) multi-attribute plots are presented, which can be considered as extended versions of those coming from PanelCheck, since they can handle much more complicated settings, such as unbalanced data and more complex mixed effects models.

In support of the methodological developments I have written the open-source and free software packaged **lmerTest** (`http://cran.r-project.org/web/packages/lmerTest/`) and **SensMixed** (`https://r-forge.r-project.org/R/?group_id=1433`) for the free software package, R (`http://www.r-project.org/`; R Development Core Team, 2011). The **lmerTest** package has developed into a more generally applicable statistical package for analyzing linear mixed effects models.

1.1 Overview of the thesis

This thesis consists of a number of papers, material related to two R packages and five chapters providing an introduction to the appended papers and the R packages. Four of the six papers included in this thesis were written for the journal of Food Quality and Preference - the main sensometrics journal. An additional two appendices contain material associated with two R packages

In the following sections the main chapters, the journal papers, and the R packages **lmerTest** and **SensMixed** are introduced, linked to each other and put into the appropriate context.

1.1.1 Main chapters

The main chapters of the thesis are intended to provide background foundation for the papers and material on the R packages in the appendices.

In Chapter 2 the theory behind the linear mixed effects models is described. The chapter mainly focuses on introducing different methods for performing inference on the fixed effects. The Sattherthwaite's method for approximation to degrees of freedom for the F and t tests for the fixed effects is described in details, since this method is implemented in the **lmerTest** package. The examples are provided where different methods are compared between each other in testing the fixed effects.

in Chapter 3 the data sets, that are typically used in sensory and consumer studies are described together with the commonly used in such studies linear mixed effects models. Then issues regarding these simple mixed effects models as discussed, which serve as a motivation for the paper in Appendix A. Finally mixed assessor models (MAM) are considered, that are able to correct for the scaling effects in the sensory data. The methods, that can detect scaling effects as well as the methods that are commonly used to correct for the scaling are described there as well. The extensions to MAM are defined in this Chapter as well together with an example showing the usefulness of such models as applied to sensory data..

Chapter **??** provides the description of the visualization tools, that are frequently used in sensory and consumer studies. The Chapter also introduces novel multi-attribute plots, that are implemented in the **SensMixed** package and introduced in paper in Appendix C.

Chapter 5 describes the software tools developed in the period of my PhD studies and that support as well the methodology developments introduced in previous chapters. Here first the **lmerTest** package is introduced and its functions are illustrated via examples. Some implementation details are provided here as well. The **lmerTest** package can be considered as a core tool for all the papers, that form the thesis as well as software tools. Then the **SensMixed** package is introduced together with some implementation details.

Chapter 6 includes concluding remarks.

1.1.1.1 Journal Papers

The first paper included in appendix A is written for and published in Food Quality and Preference. In this paper a model building approach was proposed for analysis of sensory and consumer data as well as analysis of preference mapping within a mixed effects model framework. It was shown that considering the biggest possible model at the first place and simplifying it using the stepwise selection process provides more insight into the data otherwise ignored due to the choice of too simplistic models in a variety of situations in sensory and consumer studies. The **lmerTest** package provides software support for the presented methodology.

The second paper included in appendix B is written for and submitted to Journal of Statistical Software. This paper can be considered as a so-called package vignette for the R-package **lmerTest**, which has a general statistical input, and not only constrained to sensometrics field. A novel contribution forms part of the package - the implementation of Satterthwaite's approximation to degrees of freedom for the F and t tests in tests of fixed effects in linear mixed effects models. The functions facilitating developed methodology in paper in Appendix A form also part of the package. Other miscellaneous functions facilitating analysis and tests of the linear mixed effects models are included in the package as well. The paper illustrates the usefulness of the package in a number of examples.

The third paper included in appendix C is dedicated for Food Quality and Preference Journal. In this paper first of all the methodology introduced in Kuznetsova et al. (2015) is adopted for analysis of multi-attribute sensory data. Then a couple of models are proposed for analysing sensory data, which can be considered as an extended versions of the mixed assessor model proposed by Brockhoff et al. (2015). The extension consists on handling more complex random and fixed structures as well as handling unbalanced data. A new visual tool for the analysis of multi-attributes sensory data introduced in paper in Appendix E is implemented. A user-friendly application accompanies the package, the tutorial for it can be found in Appendix I.

The fourth paper included in appendix D is intended for Journal of Statistical Software. The paper introduces ConsumerCheck software (http://consumercheck.co/). The software is a new, open-source, dedicated for analysis of consumer preference data. The ConsumerCheck software is a by-product of an international research project, which also financed my PhD studies. The aim was to develop a software that could provide an alternative software package that is open source and that has a an easy-to-use graphical user interface that makes the statistical methods available to users that have little or no programming skills. The core of the software was developed using Python software. I wrote

the conjoint analysis part of the software using the statistical language R. Conjoint analysis is a method for analyzing the effects of consumer characteristics on consumer preferences (Gustafsson et al., 2003). A common approach is to analyze it in a mixed effects model framework, where consumers are treated as randomly selected from large population of consumers. Mixed effects models are then constructed using the R-package **lme4** (Bates et al., 2013). The tests and post-hoc analysis for the models are performed using the **lmerTest** R-package (Kuznetsova et al., 2013a). For flexibility purposes, different degrees of complexity of mixed effects models as well as model selection process can be chosen by the user.

The fifth paper included in appendix E is intended for Food Quality and Preference journal. The paper presents a new approach of presenting a multi-attribute plot for sensory studies, a so-called $\tilde{d}$-plot. In this paper the effect size $\tilde{d}$ is proposed together with the method for calculating the estimates of it. The paper shows that the $\tilde{d}$ can be very useful for comparing the effect sizes of different factors and different attributed coming from the sensory studies. In the **SensMixed** the calculation of the estimates of the $\tilde{d}$ is implemented.

The sixth paper included in appendix F is written and published in Food Quality and Preference journal. In this paper different mixed effects models are constructed and compared between each other as applied to the sensory data, coming from Bavay et al. (2013). There it is shown that it is important to account for possible carry-over effects, otherwise the results lead to improper conclusions. Moreover it was pointed out, that it is also crucial to combine carry over effects with possible scaling effects.

1.1.1.2 The lmerTest package

The official description of the package (cf. the reference manual, appendix G) states: "Different kinds of tests for linear mixed effects models as implemented in the **lme4** package are provided. The tests comprise types I - III F tests for fixed effects, LR tests for random effects. The package also provides the calculation of population means for fixed factors with confidence intervals and corresponding plots. Finally the backward elimination of non-significant effects is implemented".

The **lmerTest** was originally motivated by performing conjoint analysis for preference data as part of the ConsumerCheck software in 2010. It was decided that the conjoint analysis should be written using the open-source R language and the **lme4** package for fitting mixed effects models. One of the challenges we faced was to be able to do a model-building with a step-wise approach as well as providing

the tests for the fixed as well as random effects and presenting the results in a nice, user-friendly way. First of all, no software was available at that time for the model building purposes, and secondly no p values were provided in ANOVA table in tests for the fixed effects in the **lme4** package. The first functions for the conjoint analysis included such functionalities as finding the parsimonious model by the step-down model building approach as well as providing p values with the Satterthwaite's approximation methods for the F-tests. These functions were wrapped into a package with a name **MixMod**, which appeared in CRAN in February 2012. Eventually it was decided that the **MixMod** package and its functions were useful beyond the sensometrics field and an implementation was aimed for a wider audience in the **lmerTest** package. The **lmerTest** first appeared on the comprehensive R archive network, CRAN (http://cran.r-project.org/web/packages/lmerTest/index.html)in January 2013. The code development has been hosted on R-FORGE (https://r-forge.r-project.org/R/?group_id=1433) supporting Subversion (http://subversion.apache.org/) for software versioning and revision control. Because of a great demand of getting p values in tests for fixed effects for the `lmer` objects in the R community, we have overloaded the `lmer` function as well as the `summary` and `anova` functions in providing the p values with the Satterthwaite's approximation method. The `step` method produced a stepwise selection process. In 2014 a number of functions were rewritten. The most significant user visible change was the computational time.

1.1.1.3 The SensMixed package

The official description of the **SensMixed** package (cf. the reference manual, appendix H) reads: "The package provides functions that facilitate analysis of Sensory as well as Consumer data". Two main functions are part of the package: `sensmixed` and `consmixed`. The `consmixed` is dedicated for analysis of consumer preference data in a mixed effects model framework. The functionality of the `consmixed` is in fact the same as in the conjoint analysis part of the ConsumerCheck, however we decided to include the function `consmixed` in the **SensMixed** package in order to provide one software tool where analysis of both sensory and consumer data within a mixed effects model framework can be performed. The `sensmixed` function offers mixed model ANOVA specifically prepared for multi attribute sensory data. First it offers the automated construction of mixed effects models of different complexities and finding the parsimonious models by adopting the `step` function from the **lmerTest** package. Then the new tool introduced in in paper in Appendix E to visualise the results is implemented in the package as well. Also an extended version of the mixed assessor model introduced in Brockhoff et al. (2015) is implemented in the package together with the post-hoc analysis. Last but not least the **SensMixed** package

includes a **shiny** application Chang et al. (2015), which provides an intuitive and easy-to-use graphical user-interface for the `sensmixed` and `consmixed` functions. The tutorial supporting the the application can be found in Appendix I. The tutorial is written only for the analysis of the sensory data, since the analysis of the consumer data simply replicates the conjoint analysis part of the ConsumerCheck software. The **SensMixed** is hosted on the comprehensive R archive network, CRAN (https://cran.r-project.org/web//packages/SensMixed/index.html) since August 2015.

Chapter 2

Linear mixed effects models (LMM)

In 1918 Ronald Fisher introduced random effects models in his paper Fisher (1918). Thereafter mixed modeling, where both random and fixed effects are modelled, has become a great area of statistical research. Mixed effects models are prominently used in research involving human and subjects in fields such as biology, ecology, marketing and others, and have also been used in industrial statistics. Throughout the thesis I consider linear mixed effects models, which model the fixed and random effects as having a linear form and where the response variable is assumed coming from the Gaussian distribution. Generalized linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as e.g. binary responses. In this Chapter I am going to shortly introduce the theory behind the linear mixed effects models, then I will discuss different methods for the inference on parameters of LMM. This Chapter can be considered as a theoretical basis for the thesis.

2.1 Outline of LMM

The standard linear model, which is one of the most common statistical models is:

$$y = X\beta + \epsilon \quad \epsilon \sim N_n(0, R) \tag{2.1}$$

Here y represents a vector for observed data, β an unknown vector of fixed-effects parameters with known design matrix X and ϵ is unknown random vector. The mixed models is a generalization of the standard linear model in a way that an additional variation in the data may be accounted for:

$$y = X\beta + Zu + \epsilon \quad u \sim N_k(0, G) \quad \epsilon \sim N_n(0, R) \tag{2.2}$$

with β representing all fixed-effects parameters, u the random-effects, X the $n \times p$ design matrix for the fixed-effects parameters, and Z the $n \times k$ design matrix for the random-effects. The matrix Z can contain either continuous or dummy variables, similar to the design matrix X. Throughout the thesis the focus will be on a simplified version of the model in Equation (2.2), where u and ϵ are independet and $R = \sigma^2 I$. Then the covariance matrix of y in Equation(2.2) would be:

$$\begin{aligned} V &= var(y) = var(X\beta + Zu + \epsilon) \\ &= var(X\beta) + var(Zu) + var(\epsilon) \\ &= var(Zu) + \sigma^2 I = Z^\top G Z + \sigma^2 \end{aligned} \tag{2.3}$$

The likelihood function is a function of the observations and the model parameters. It returns a measure of the probability of observing a particular observation y, given a set of model parameters β and θ. Here θ is the vector of parameters used in the two covariance matrices G and R. Frequently, due to the convenience, the negative log-likelihood l is considered instead:

$$l(y, \beta, \theta) = \frac{1}{2} n log(2\pi) + log|V(y)| + (y - X\beta)'(V(\theta))^{-1}(y - X\beta) \tag{2.4}$$

A method, that is often used for estimating model parameters is the maximum likelihood method. where the parameter estimates are found in the following way:

$$(\hat{\beta}, \hat{\theta}) = \underset{(\beta,\gamma)}{\arg\min}\, l(y, \beta, \theta) \tag{2.5}$$

However the maximum likelihood method tends to underestimate the random effects parameters. A modification then is generally considered, known as

restricted maximum likelihood method (REML), where instead of l in Equation (2.4) the following version is considered :

$$l_{REML} = \frac{1}{2} nlog(2\pi) + log|V(y)| + (y - X\beta)'(V(\theta))^{-1}(y - X\beta) + log|X'(V(\theta))^{-1}X| \tag{2.6}$$

The REML random effects are then not biased, at least is balanced cases.

2.2 Inference on random effects

Likelihood ratio test (LRT) is a commonly used test on random effects. In this test statistic $T = 2(l - l_0)$ is considered, where l and l_0 represent the log-likelihoods of two nested models, which have the same fixed structure. T follows asymptotically a χ^2 distribution with degrees of freedom being the difference between number of papameters in the two models. However the p values coming from this test can be conservative. This is due to the fact that changing from the more general model to the more specific model involves setting the variance of certain components of the random-effects to zero, which is on the boundary of the parameter region, hence asymptotic results for LRT have to be adjusted for boundary conditions. Following Self and Liang (1987); Stram and LEE (1994) the LRT more closely follows an equal mixture of χ^2-distributions with zero degrees of freedom (a point mass distribution) and one degree of freedom. The p-value from this test can be obtained by halving the p-value.

2.3 Inference on fixed effects

Similarily to the tests on random effects likelihood ratio test can be used to test the fixed part of a mixed model. The LRT statistic is:

$$T = 2(ll - ll_0) \tag{2.7}$$

where ll and ll_0 represent the log-likelihoods of two nested models with the same random structure. T follows asymptotically a χ^2 distribution. As described in Pinheiro and Bates (2000) a likelihood ratio test for models based on REML is not feasible, because there the last term in the REML criterion in Equation 2.6 changes with the change in the fixed-effects specification. Subsequently, LMM models with different fixed-effects structures fit using REML cannot be

compared on the basis of their restricted likelihoods. Thus the ML model fits should be used in LRT tests. Even though a likelihood ratio test for the ML fits of models with different fixed effects can be calculated, one should use the test with caution, since it can produce anti-conservative p-values, as also pointed out in Pinheiro and Bates (2000).

One may consider an F test of the hypothesis $H_0 : L\beta = 0$, where L is a contrast matrix of $q = rank(L) > 1$. A test statistic for this hypothesis is:

$$F = \frac{(L\hat{\beta})^\top (L\hat{C}L^\top)^{-1}(L\hat{\beta})}{q} \tag{2.8}$$

A one-dimensional case of the F test corresponds to a t test. For a hypothesis $H_0 : l\beta = 0$, where l is a vector the t statistics is:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{2.9}$$

Where $\hat{C}$ is an estimated variance-covariance matrix. For complex situations, such as, for example, unbalanced data sets, the F and t test statistics follow unknown distributions. Approximation methods might then be applied, that assume that the test statistics approximately follow an F-distribution with carefully calculated degrees of freedom. The widely used methods are the Satterthwaite's based and Kenward-Roger's methods of approximations which I discuss in the following sections.

2.3.1 Satterthwaite's approximation

To find the proper degrees of freedom df of the t distribution a proper χ^2 distribution needs to be found for the variance estimate $\hat{\text{var}}(l\hat{\beta}) = l\hat{C}l^\top$. Giesbrecht and Burns (1985) followed Satterthwaite (1946) idea, where the degrees of freedom df are obtained such that the relationship between mean and variance of the statistic, which distribution we are trying to approximate, is the same as for the χ^2 distribution:

$$\text{Variance} = \frac{2(\text{mean})^2}{df}$$

from which the degrees of freedom df may be obtained:

$$df = \frac{2(l^\top \hat{C} l)^2}{[var(l^\top \hat{C} l)]}$$

Taking $f(\theta) = l^\top C(\theta) l$, $var(f(\theta))$ can be approximated by applying univariate delta method as:

$$var(f(\theta)) \approx [\nabla_{f(\theta)} \hat{\theta}]^\top A [\nabla_{f(\theta)} \hat{\theta}] \tag{2.10}$$

where $\nabla_{f(\theta)} \hat{\theta}$ is a vector of partial derivatives of $f(\theta)$ with respect to θ evaluated at $\hat{\theta}$. A is the variance covariance matrix of the $\hat{\theta}$-vector, which can be found from the second derivatives of the log-likelihood function.

In a multi-degree-of-freedom F test the idea is to first eigen-decompose the matrix $(L\hat{C}L^\top)^{-1}$ so that qF can be written as a sum of squared t-variables:

$$qF = \sum_{m=1}^{q} t^2_{df_m}$$

then, following Giesbrecht and Burns (1985), the Satterthwaite method-of-moment is applied in order to get the approximations to degrees of freedom df_m for each of the t_{df_m}, which are regarded to be approximately distributed as Student's t distribution, Finally, by using the followong equations:

$$\sum_{m=1}^{q} E(t^2_{df_m}) = \sum_{m=1}^{q} \frac{df_m}{df_m - 2} = E_Q$$

$$E(F_{q,df}) = \frac{df}{df - 2} \text{ for } df > 2$$

, the denominator degrees of freedom can be calculated:

$$df = \frac{2E_Q}{E_Q - q}$$

2.3.2 Kenward-Roger's approximation

In Kenward-Roger's method the estimated variance covariance matrix $\hat{C}$ in Equation 2.8 is adjusted in order to improve the small sample distributional

properties of F and then the denominater degrees are found via Satterthwaite's based method-of-moment. For the details regarding the method I refer to Kenward and Roger (1997) and Halekoh and Højsgaard (2014a). Generally the Kenward-Roger's method outperforms the Satterthwaite's, especially in a small sample unbalanced situations (Schaalje et al., 2002). However, in a great randge of situations they provide the p values, which are in the same order of magnitude (I will discuss that in the following Sections). The advantage of the Satterthwaite's method is that it is generally faster. Some comparisons are made in paper in Appendix B between the computational time for Satterthwaite's and Kenward-Roger's methods.

2.3.3 Comparisons of F tests with the LRT

In the paper in appendix B a simulation study is presented, showing the clear superiority of the F tests with both approximation methods compared to the LRT test. Halekoh and Højsgaard (2014b) presented an example, where they compared F test with Kenward-Roger's approximation to the LRT applied to the Mississipi influents data, which comes from the **SASmixed** package by Littell et al. (2014). Here I use the same data and example in order to also compare with the Satterthwaite's approximation method. The Mississippi data contains the nitrogen concentration at six randomly selected influents of the Mississippi river from the following three types of sites:

Type 1	Type 2	Type 3
No farm land in watersheld	Less than 50% farm land in watershed	More than 50% farm land in watershed
Influents 3 and 5	Influents 1, 2, and 4	Influent 6

Since not every level of `Type` appears with every level of `influent`, the data is unbalanced. The main interest in this data is on determining how much of the variation in the influents can be attributed to the different levels of `Type`. It is reasonable to consider `Type` as a fixed effect and `influent` as a random effect resulting in the following mixed effects model:

$$y_i = Type_i + u(Type_i) + \epsilon_i$$

with $u_j \sim N(0, \sigma^2_{influents})$ and $\epsilon \sim N(0, \sigma^2)$

The Table 2.1 represents the *LRT* test for the test of the `Type` effect. From the Table it is seen that the *LRT* test suggests a highly significant effect of

Type. However, since there are not many observations in the data, trusting large-sample asyptotic results is questionable (Pinheiro and Bates, 2000). Note that here the ML was used in fitting the nested models for the reasons discussed in Section 2.3.

Table 2.1: LRT for the fixed-effect Type. Generated by lme4 package

	Df	AIC	deviance	Chisq	Chi Df	Pr(>Chisq)
..1	3	262.56	256.56			
object	5	256.57	246.57	9.98	2	0.00679

The ANOVA table with the F test for the fixed effect Type with the Kenward-Roger's approximation method is presented in Table 2.2.

Table 2.2: F-test for the fixed effect Type with Kenward-Roger's approximation to degrees of freedom. Table generated by lmerTest. Kenward-Roger's method generated by pbkrtest.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
Type	541.55	270.77	2	3.32	6.37	0.0731

The ANOVA table with the F test for the fixed effect Type with the Satterthwaite's approximation method is presented in Table 2.3.

Table 2.3: F-tests for the fixed effect Type with Satterthwaite approximation to degrees of freedom. Generated by lmerTest

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
Type	541.76	270.88	2	3.39	6.37	0.071

From Tables 2.3 and 2.2 it can be observed that the p values coming from the approximation methods are very close to each other and claim that the `Type` effect is non-significant according to the 0.05 Type 1 error rate. The following analysis of means supports the results of Kenward-Roger's and Satterthwaite's methods:

Table 2.4: F-tests for the fixed effect Type from a linear regression analysis, where y is averaged across influents

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	2	298.28	149.14	7.07	0.0732
Residuals	3	63.28	21.09		NA

Table 2.4 represents the ANOVA table for a linear regression, where the response variable is a mean value of concentration of nitrogen according to both `influents` and `Type` factors and the dependent variable is the factor `Type`. This example shows that the LRT test can produce very different p values than the F test with approximation methods. For small sample and unbalanced data the p values from the LRT can indeed be anti-conservative. The Satterthwaite's method produces very similar results to Kenward-Roger's even if the data here is unbalanced and the number of observations in the data is not high.

In the generation of the Tables 2.2 and 2.3 the **lme4** package was used (Bates et al., 2013) for fitting LMM and the **lmerTest** for generating ANOVA tables. The Kenward-Roger's method was used through the `KRmodkomp` function of **pbkrtest** package. The Satterthwaite's method was used through the **lmerTest** package, which I describe in Chapter 5.1. Calculating the means across factor `influent` used the `summaryBy` function from the **doBy** package (Højsgaard et al., 2014a).

2.3.3.1 Random coefficient regression - A simulation study

In this example I perform a simulation study, which is inspired by an example coming from Halekoh and Højsgaard (2014a), where the authors performed a simulation study in order to compare LRT to Kenward-Roger's and bootstraping methods. Here, I make comparisons between Satterthwaite's, Kenward-Roger's methods and the LRT. This simulation study was actually originally performed by Kenward and Roger (1997) in purposes of investigating the Kenward-Roger's method. The following random coefficient model was considered there:

$$y_{jt_j} = \beta_0 + \beta_1 t_j + A_j + B_j t_j + \epsilon_{jt_j}$$

with

$$Cov(A_j, B_j) = \begin{bmatrix} 0.250 & -0.133 \\ -0.133 & 0.250 \end{bmatrix} \text{ and } Var(\epsilon_{jt}) = 0.25 \qquad (2.11)$$

where there are $j = 1, \ldots, 24$ observed subjects divided into three groups of eight subjects. For each group observations are made at the non overlapping times $t = 0, 1, 2; t = 3, 4, 5$ and $t = 6, 7, 8$. The data for the simulation were generated under the assuption that $\beta_0 = \beta_1 = 0$, (A_j, B_j) and ϵ_{jt} are normally distributed with zero expectation, (A_j, B_j) are independent from ϵ_{tj} and observations from different subjects are independent.

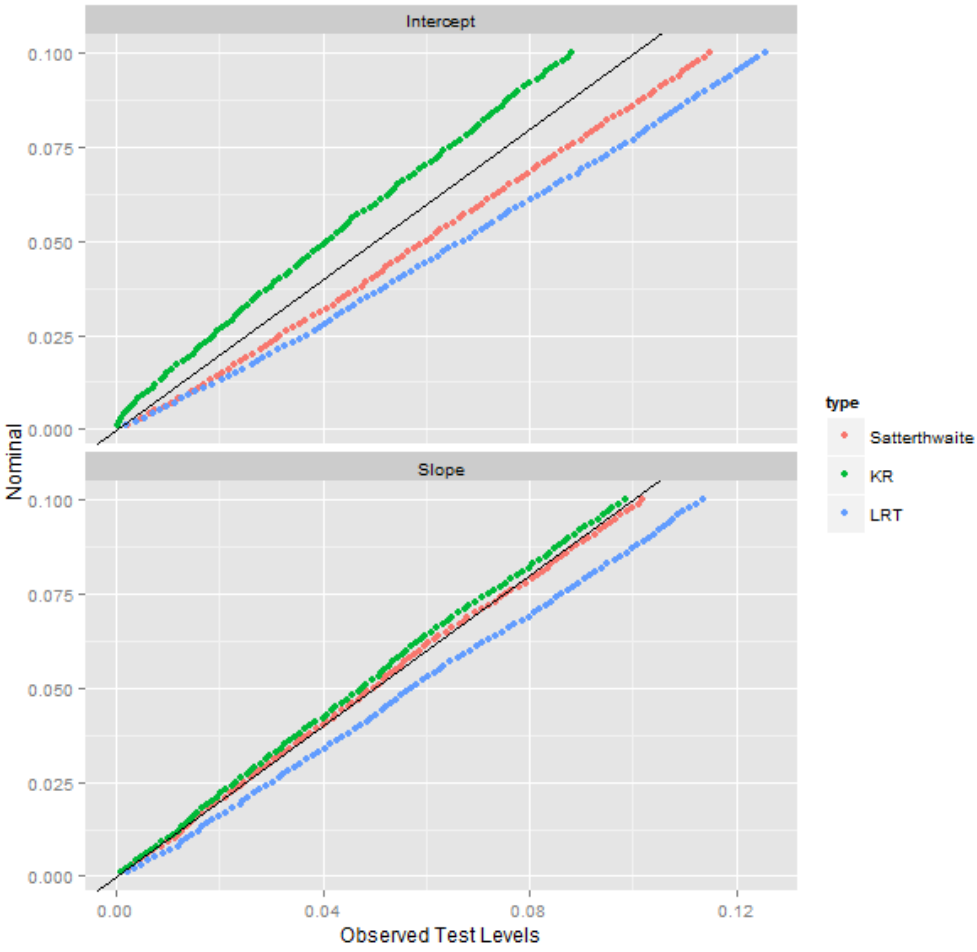


Figure 2.1: Empirical p values versus nominal p values ranging from 0.001 to 0.12 for the test of the presence of the slope and intercept fixed effects. The results are based on 20000 simulations

The results of the simulations are presented in Figure 2.1. The LRT test gives for all nominal levels anti-conservative p-values. The observed test levels for the slope for Satterthwaite's as well as for Kenward-Roger's methods are very close to the nominal levels, the Satterthwaite's are even somewhat closer. In test for the intercept the Kenward-Roger's p values are slightly conservative whereas the p-values coming from the Satterthwaite are slightly anti-conservative. This

example shows that indeed, as is also discussed in Schaalje et al. (2002), the Kenward-Roger's method outperforms Satterthwaite's (at least in test for the intercept effect). However, the p values are within the same order of magitude. Both approximation methods outperform LRT.

2.4 Types of hypothesis testing

The key step in constructing the F test concerning fixed effects is on specifying the hypothesis contrast matrix L in Equation 2.8. Three types of hypothesis were introduced by SAS SAS (1978), but are now widespread and most of commercial as well as open source software can produce them. These types of hypotheses are introduced there in terms of sums of squares. In order to illustrate them, I consider a hypothetical data with two factors A and B and the following linear model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{2.12}$$

where α is a main fixed effect for factor A, β is a main fixed effect for factor B and γ is an effect standing for the interaction between A and B. $SS(\alpha, \beta, \alpha\beta)$ denotes sums of squares for a full model with two main effects and interaction between them. Similarly $SS(\alpha, \beta)$ denotes sums of squares for a model without the interaction effect, $SS(\alpha\beta, \beta)$ denotes sums of squares for a model that does not include main effect α. Then SAS (1978) specify the reductions of sums of squares. For instance, the following one

$$R(\alpha\beta|\alpha, \beta) = SS(\alpha, \beta, \alpha\beta) - SS(\alpha, \beta)$$

means sums of squares for interaction adjusted for the main effects.

Following SAS (1978), the types of sums of squares are then defined as:

Table 2.5: Three types of sums of squares for the effects in model in Equation 2.12

effects	Type I	Type II	Type III
α	$R(\alpha)$	$R(\alpha\|\beta)$	$R(\alpha\|\beta, \gamma)$
β	$R(\beta\|\alpha)$	$R(\beta\|\alpha)$	$R(\beta\|\alpha, \gamma)$
γ	$R(\gamma\|\beta, \alpha)$	$R(\gamma\|\beta, \alpha)$	$R(\gamma\|\beta, \alpha)$

By examining Table 2.5 it can be observed that the Type I performs the sequential decomposition of the contributions of the fixed-effects. As can be also seen

the Type II test for each main effect after the other main effect. Note that no significant interaction is assumed in Type II in tests for the main effects. As can be seen, the Type II is equivalent to running the Type I analysis with different orders of the factors, and taking the appropriate output (the second, where one main effect is run after the other). It can be also seen that the Types II and III do not depend on the order the effects are entered in a model compared to Type I. Following Searle (1987) the associated hypotheses tests for the model in Equation 2.12 are specified in Table 2.6 .

Table 2.6: Three types hypotheses for the model in Equation 2.12

	effects	Associated Hypothesis
Type I	α	$\alpha_i + \sum_j n_{ij}(\beta_j + \gamma_{ij})/n_{i\cdot}$ equal $\forall i$
	β	$\sum_i n_{ij}(\beta_j + \gamma_{ij}) = \sum_i \sum_s \frac{n_{ij}n_{is}}{n_{i\cdot}}(\beta_s + \gamma_{is}) \quad \forall j$
	γ	$\gamma_{ij} - \gamma_{ij'} - \gamma_{i'j} + \gamma_{i'j'} = 0 \quad \forall i, i', j, j'$
Type II	α	$\sum_j n_{ij}(\alpha_i + \gamma_{ij}) = \sum_j \sum_s \frac{n_{ij}n_{js}}{n_{\cdot j}}(\alpha_s + \gamma_{sj}) \quad \forall i$
	β	$\sum_i n_{ij}(\beta_j + \gamma_{ij}) = \sum_i \sum_s \frac{n_{ij}n_{is}}{n_{i\cdot}}(\beta_s + \gamma_{is}) \quad \forall j$
	γ	$\gamma_{ij} - \gamma_{ij'} - \gamma_{i'j} + \gamma_{i'j'} = 0 \quad \forall i, i', j, j'$
Type III	α	$\alpha_i + \bar{\beta}_\cdot + \bar{\gamma}_{i\cdot} = \alpha_{i'} + \bar{\beta}_\cdot + \bar{\gamma}_{i'\cdot} \quad \forall i, i'$
	β	$\beta_j + \bar{\alpha}_\cdot + \bar{\gamma}_{\cdot j} = \beta_{j'} + \bar{\alpha}_\cdot + \bar{\gamma}_{\cdot j'} \quad \forall j, j'$
	γ	$\gamma_{ij} - \gamma_{ij'} - \gamma_{i'j} + \gamma_{i'j'} = 0 \quad \forall i, i', j, j'$

From Table 2.6 we can see that the hypotheses for the interaction effect are the same for all types. It can be as well seen that in tests for the main effects the Type I and II hypotheses depend on the cell frequences and, hence become hard to interpret. On the contrary the Type III hypotheses for the main effects do not depend on the cell frequences and test the effect of one factor when averaged over all levels of the other factor. When the data is balanced, that is when all n_{ij} are equal, the three types of hypotheses become the same.

Which type of hypothesis to use has led to an ongoing controversy in the field of statistics (Speed et al., 1978; Senn, 2007; Langsrud, 2003; Macnaughton, 2009). However, it essentially comes down to testing different hypotheses about the data. Types II and I hypotheses are not easy to understand in unbalanced situations as they represent comparisons of weighted averages with the weights being a function of the cell frequencies. The Type III hypotheses are easy to interpret. Type II are natural to consider when the cell frequencies are somehow indicative of population characteristics, Type I is appropriate when a logical ordering exists for the effects being examined Speed et al. (1978). Another situation, where the Type I hypothesis test is preferable, is a mixed assessor

model, that will be discussed in the Chapter 3.6.

The Type III hypothesis test can be valuable in studies coming from the consumer science. (Macnaughton, 2009) gives a nice example from the marketing field, where the usefullness of the Type III hypothesis test is illustrated.

2.5 Post-hoc analysis

Post hoc testing becomes important when one of the tests for the fixed effects shows significanct effect (Wiley, 1962). In consumer and sensory studies, for example, if one finds that the product effect is significant, one will be interested in knowing more about which products that are different from each other. Are all of them different or is it just a clear difference between two of the products? In such cases one can compute the averages for each of the products and compare them visually, but it is in general useful also to accompany this check with a statistical testing procedure. One may consider a t test of the hypothesis $H_0 : l\beta = 0$, where l is an appropriate contrast vector. A test statistic for this hypothesis is:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{2.13}$$

where $\hat{\beta}$ are the estimates of the fixed effects, $\hat{C}$ is the estimate of variance-covariance matrix of fixed effects. As in the case of the F test in Section 2.3, the t statistics does not follow exactly the t distribution, only in balanced situations it does. So the degrees of freedom can not be directly calculated. Similarily to the F test the methods to approximation to degrees of freedom can be applied.

The confidence intervals for the pairwise comparisons can be then obtained via the following equation:

$$CI = l\hat{\beta} \pm t_{\frac{\alpha}{2}}(\nu) * \sqrt{l\hat{C}l^\top} \tag{2.14}$$

where α is a Type 1 error rate (commonly it is 0.05).

Comparing levels of a factor in question (e.g. product effect in sensory/consumer studies) using t tests can involve many tests, a possible risk is that too many tests become significant. In order to control this type of error a number of methods, the multiple testing corrections , have been developed that control the overall significance level. The commonly used one is the Tukey method Tukey (1949). Other methods include Bonferroni Dunn (1961), Newman-Keul's Steel

et al. (1997) and others. The Bonferroni method is very easy in implementation, however it can be conservative.

Chapter 3

LMM in Sensory and Consumer studies

This Chapter consists of two parts. In the first one I introduce the types of the studies that are common in sensometrics field. Then I decribe the LMMs, that are commonly considered for analysis of such studies. Finally I introduce the automated approach of analyzing the studies within a mixed effects model framework suggested in paper in Appendix A. In the second part the mixed assessor models proposed by Brockhoff et al. (2015) are considered. I introduce their extended versions and their analysis within an automated approach.

3.1 Typical studies

3.1.1 Sensory studies

Sensory profiling is one of the most used methods in sensory analysis (Lawless and Heymann, 2010). The method consists on describing differences between products by trained sensory assessors, so-called sensory panel. Figure 4.4 represents the structure of a typical sensory data. Generally, a sensory panel consists of 10-15 assessors. These assessors are trained to detect small differences between products according to some prespecified characteristics of the products,

attributes. Typically the number of attributes is between 10 - 15. The assessors are then asked to put scores to the products for the selected attributes. There are different scale ranges for the scores, but the common ones are 1-7, 1-12, where "1" means low intensity and "12" means high intensity.

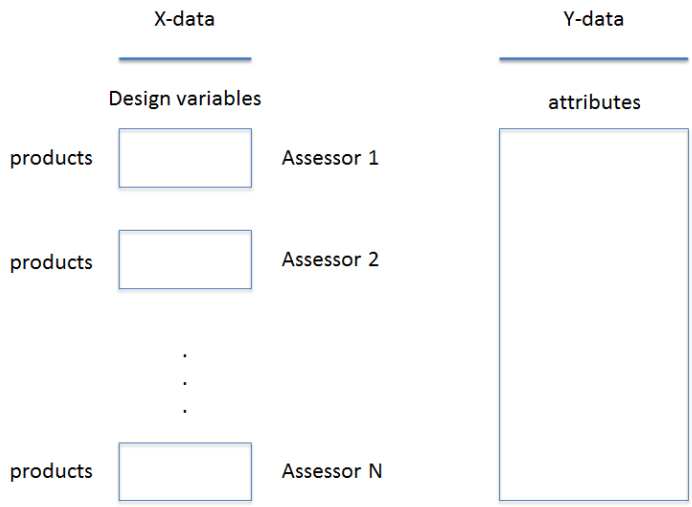


Figure 3.1: structure of a typical sensory data set

3.1.1.1 Example: TVbo data

In this Chapter and in the following ones the `TVbo` data set will be used in examples. In order not to repeat the description of the data multiple times, I introduce the data here. The `TVbo` data was produced by the highend HIFI company Bang and Olufsen A/S, Struer, Denmark, and was used for a workshop at the 8th Sensometrics Meeting in 2008 https://www.compusense.com/sensometrics2008/. In this data the main purpose was to assess 12 products, specified by two features: Picture (factor with 4 levels) and TVset (factor with 3 levels). The products were assessed by 8 assessors in 2 replications for 15 different attributes.

3.1.2 Consumer studies

The aim of the sensory studies is to decribe the products as objectively as possible. In order to obtain information regarding what people like, different types of consumer studies are needed. In these studies generally consumers are selected randomly from a certain population. In most cases the number of consumers is between 100 - 150. The consumers are asked to score the products in terms of their liking, preference or purchase intent (Lawless and Heymann, 2010; Naes et al., 2010). A typical structure of the consumer data is presented in Figure 3.2. There are different types of consumer studies. Consumer studies with a hedonic response are typical and largely used both in industrial and in a research context. Here the consumers are asked to put scores to the products on a scale ranging from 1 to 7, or 1 to 9. Other types of consumer studies include ranking tests, where products are presented simultaneously and the consumers are asked to rank them according to liking/purchase intent. Choice tests are also frequently used, where the consumers are given a number of so-called choice sets and for each choice set they are asked to select the product they like/prefer the best. Other methods of measuring affective responses can be found in MacFie (2007). In the thesis consumer studies with a hedonic response will only be considered. In consumer studies one would generally be also interested in understanding individual differences between consumers in a better way. Therefore frequently data containing consumer characteristics such as gender, age e t.c. is collected as well.

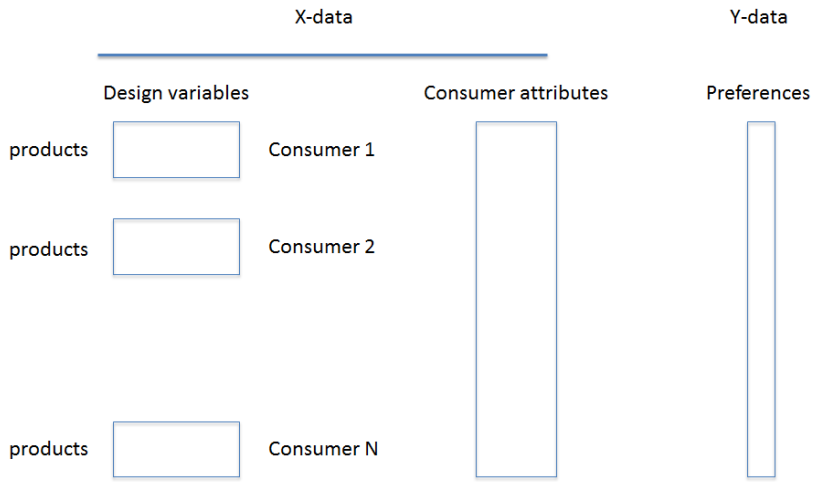


Figure 3.2: structure of a typical consumer data set

3.1.3 Preference mapping

Preference mapping is a techique that links two types of analysis: consumer and sensory (Lawless and Heymann, 2010; Naes et al., 2010). The techique is of particular importance since it can eventually detect what are the main drivers of liking. For example, is the preference of a certain product related to sweetness or other sensory attribute? Since usually the attributes from the sensory study are correlated, a common technique is to apply principal component analysis (PCA) and extract first few principal components, that contain most of the variation in the data. Then the classical external preference mapping technique consists on fitting regression model for the preference as a function of the extracted PCA components for each individual consumer. THis is a so-called external preference mapping, which I consider in the thesis. Other version of the preference mapping is the internal preference mapping, where first the PCA of the consumer liking data is performed, and then the correlation coefficients of the sensory variables with the PC scores are calculated (MacFie, 2007). Sometimes consumers are clustered first, and then the preference mapping is applied to each cluster.

3.2 LMM and sensory studies

Let us consider a simple example of a sensory experiment where we have I assessors, J products and R replicates. This type of data can be described by a mixed ANOVA for replicated two-way data Naes et al. (2010), where as effects we have factors A (assessor) and B (product). A reasonable model can then be written as

$$y_{ijr} = \mu + a_i + \beta_j + d_{ij} + \epsilon_{ijr} \tag{3.1}$$

where a_i and β_j are main effects for factors A and B and d_{ij} is the effect corresponding to interaction between A and B. If we consider the effects of factor A random, then this implies that the effects a_i (assessor) and d_{ij} (interaction between assessor and product) are random:

$$\begin{aligned} a_i &\sim N(0, \sigma^2_{assessor}) \\ d_{ij} &\sim N(0, \sigma^2_{assessor \times product}) \\ \epsilon_{ijr} &\sim N(0, \sigma^2_{error}) \end{aligned} \tag{3.2}$$

the model in Equation (3.1) can be written on the matrix form in Equation (2.2)

with:

$$
\begin{aligned}
\beta &= (\mu, \beta_1, \ldots, \beta_J) \\
u &= (a_1, \ldots, a_I, d_{11}, \ldots, d_{IJ}) \\
n &= R \cdot J \cdot I \\
p &= 1 + J \\
k &= I + I \cdot J \\
G &= \begin{pmatrix} \sigma^2_{assessor} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{assessor \times product} \end{pmatrix} \\
R &= \begin{pmatrix} \sigma^2_{error} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{error} \end{pmatrix}
\end{aligned}
\tag{3.3}
$$

Here matrix G has a so-called variance components structure. In sensory studies the main interest is in testing the product effect. Whenever data is balanced, the F statistics for the test of product effect can be calculated using the following formula:

$$F = \frac{MS_{product}}{MS_{product \times assessor}}$$

which follows exactly an F distribution with $(J-1, (I-1) \times (J-1))$degrees of freedom. If the data is unbalanced, then the distribution of F is unknown. Different approximation methods are available, which assume F follows an F distribution and the denominator degrees of freedom are approximated. Two of such methods I discussed in Chapter 2 in Sections 2.3.1 and 2.3.2.

3.3 LMM and consumer studies

A so-called joint ANOVA approach, where consumers are treated as random effects, leads to the use of the mixed effects models (Naes et al., 2010). Let us for illustration consider a simple example where there are:

- N consumers
- J products

- K background information on consumers (e.g. gender)

Then the following mixed effects models can be considered:

$$y_{jkn} = \mu + \beta_j + \gamma_k + \beta\gamma_{jk} + (Cons \times \beta)_{jn} + (Cons \times \gamma)_{kn} + \epsilon_{jkn} \tag{3.4}$$

where $Cons$ stands for a random consumer effect, β stands for a fixed product effect and γ stands for a fixed gender effect.

$$\begin{aligned} (Cons \times \gamma)_{kn} &\sim N(0, \sigma^2_{consumer(gender)}) \\ (Cons \times \beta)_{jn} &\sim N(0, \sigma^2_{consumer \times product}) \\ \epsilon_{jkn} &\sim N(0, \sigma^2_{error}) \end{aligned} \tag{3.5}$$

3.4 LMM and preference mapping

The most commonly used preference mapping methods are those that are based on a linear model for the relationship between the sensory and preference data for each consumer. Let us consider a simple example, where a sensory as well as consumer data are available. The principal component analysis is first applied and the first two principal components are extracted. A common approach in the external preference mapping is to relate the sensory scores to the acceptance data for each consumer, using the linear model Lawless and Heymann (2010); Naes et al. (2010) in a following way:

$$y_n = \beta_0 + \beta_{1n} sens_1 + \beta_{2n} sens_2 + \epsilon_n$$

where y's are the liking values, $sens_1$ and $sens_2$ are the principal components from the sensory data.

Another way is to also include quadratic term $sens_1 sens_2$ resulting in following models for each assessor:

$$y_n = \beta_0 + \beta_{1n} sens_1 + \beta_{2n} sens_2 + \beta_{3n} sens_1 sens2 + \epsilon_n$$

Yet another approach, is to relate consumer acceptance data to the sensory scores in one model framework, where consumers are treated as random effects (Naes et al., 2010). Then a mixed effects model comes into play and the following model may be considered:

$$y_n = \beta_0 + \beta_1 sens_1 + \beta_2 sens_2 + Cons_n + b_{1n} sens_1 + b_{2n} sens_2 + \epsilon_n \tag{3.6}$$

This is a random coefficient model, where a possible random structure is the following one:

$$(b_0, b_1, b_2) \sim N(0, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}), \quad Cons \sim N(0, \sigma_c^2), \quad \epsilon_n \sim N(0, \sigma^2) \tag{3.7}$$

3.5 Motivation for the automated analysis

The commonly used models in sensory and consumer studies described in this Chapter can not cover all situations that arise in sensory and / or consumer studies. For instance, in model in Equation 3.1 it is assumed that the replicates are randomized within the full experiment, which means that there is no systematic replicate effect. Such random replicate situation is very common in sensory studies, however there are other possibilities, where the same products are served in several seperate testing sessions Naes et al. (2010). Other effects might be present in the sensory studies, such as carry-over over effects. These types of effects are encountered when the studied products are prone to biological heterogeneity (e.g. fruits, vegetables etc.) (Bavay et al., 2013)

In sensory as well as consumer studies it is not uncommon that the products investigated are constructed using some kind of experimental design (Jaeger et al., 2013; Beck et al., 2014). In the `TVbo` data, for example, the 12 products are really stemming from 3×4 full factorial. In the paper in Appendix C other data of such a multi-way product stucture are presented. In practice, accounting for the multi-way product structure in the analysis amounts to decomposing product factor into sub-factors (Naes et al., 2010). In the `TVbo` data, for instance, this amounts to consider 2 main effects corresponding to Picture and TVset features and an interaction effect. The advantage of decomposition is that both the average effects of treatment factors can be investigates as well as their interaction.

In preference mapping it can be valuable to combine the sensory properties of the products with the additional background information on consumers calling for considering extra context variables, such as gender e t.c. in models such as in Equation 3.6. This calls for a need to apply ANCOVA models (ANCOVA, see e.g. Weisberg (1985)). However, the combination of sensory properties with, for instance, categorical variables can complicate not only the analysis of the models, but also the interpretation of the results. As a consequence, generally it is not advised to consider ANCOVA models in sensory and consumer studies (Naes et al., 2010).

All that calls for a need to being able to consider and analyse more complex structure in mixed effects models - both in fixed and random parts. However the question arises as to which model to consider - generally one would prefer to consider the most parsimoneous one, that would capture all the important information in the data. In paper in Appendix A the methodology is proposed, where the biggest possible model is considered at the first place, which accounts for potential carry-over effects, multi-way product structure and others. Then the parsimoneous one is found by applying the step-wise selection process. The tool supporting the methodology is implemented in the **lmerTest**, which is introduced in paper in Appendeces B and in reference manual in G. A number of examples are provided in the paper illustrating the proposed methodology.

3.6 Mixed assessor models and extensions

In sensory studies there always exist individual differences between assessors in their way of assessing the products. Some of the differences are related to the assessors sensitivity, others can be related, for example, to the use of the intencity scale. Following Naes et al. (2010), the most important types of individual differences of assessors are the following ones:

1. **Agreement** Disagreement in ranking of products
2. **Repeatability** Differences between independent replicates
3. **Discrimination** Differences in ability to discriminate between products
4. **Use of scale** Differences in mean and variability/range of the scores

The first point is related to agreement among assessors regarding the definition of the attribute in question. The second point is related to the assessors' ability to repeat a similar intensity value for the same attribute. The third point focuses on detecting differences between the products. The last point is related to how the assessors use the intensity scale. Several papers have discussed individual differences in sensory profiling and how they can be handled (for an overview, see Naes (1990); Schlich (1996); Brockhoff and Skovgaard (1994); Brockhoff (1998); Romano et al. (2008); Peltier et al. (2014); Tomic et al. (2009); Brockhoff et al. (2015)). For example, points **2 (Repeatability)** and **3 (Discrimination)** are handled by models considered in 3.2, 3.3 and 3.4. The focus of this chapter is on handling point **4 (Use of scale)** and **1. (Agreement)**, which is closely connected with point **4**. The differences on the use of the scale between assessors are generally considered to be part of nuisance effects (Naes, 1990; Tomic et al.,

2013) and should be therefore accounted for in data analysis, if it is not possible to perform an extending training for reducing them.

3.6.1 Scaling effect

Figure 3.3 illustrates individual differences in use of the scale. In the "level effect" two different assessors use the lower and the upper part of the scale. In the "range effect" assessors use the range differently. The "variability effect" shows two different assessors with different replicate error.

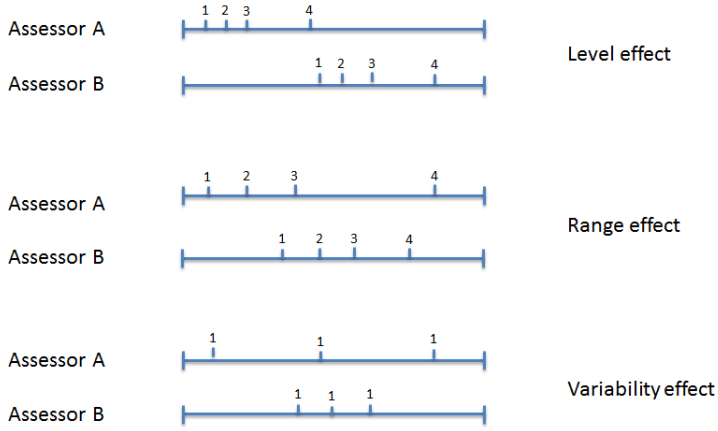


Figure 3.3: The main types of differences in use of scale

Although different use of scale is not directly related to the quality of the assessors, it is important to being able to detect it and account for it in the modelling analysis if possible (Naes et al., 2010). If individual differences in performance are simply ignored, the final results may lead to improper conclusions. Detecting scaling effects may be also helpful for obtaining better calibrated assessors for the studies in future.

3.6.2 Methods for detecting scaling effects

So called simple plots introduced in Tomic et al. (2009) and implemented in PanelCheck are useful in detecting individual differences. In correlation plot

product scores of the assessor in question are plotted versus the average scores of all assessors in the study. The plot visualizes whether an assessor rates the tested products over, under or at the same level as the panel. Another simple plot is the profile plot, which is generated for each attribute. Here the horizontal axis represents the products and the vertical axis represents the intensity scores. This type of plot can also be used in detecting level and range differences between assessors. Other methods are eggshell plot, individual line plots, consonance analysis and procrustes analysis (for an overview see Naes et al. (2010)). Another tool, that was recently proposed by Peltier et al. (2014), the so-called, MAM-CAP table, where assessor performances based on the mixed assessor models can be monitored, and where important information on each assessors performance can be obtained via compact table. The mixed assessor modelling (MAM) will be discussed in following sections.

3.6.3 Methods for handling scaling effects

The simplest pre-processing method for removing range and level effects is the standartization, where variable corresponding to an attribute in question for each assessor is scaled to unit variance and zero average. Let $\bar{y_i}^k$ and s_i^k be the mean value and the standard deviation for attribute k. The standardized measurements z_{ijm}^k can be written as:

$$z_{ijm}^k = \frac{y_{ijm}^k - \bar{y_i}^k}{s_i^k}$$

The method "makes" all assessors agree on the use of the scale. Another method, which is more sophisticated, is the method proposed in (Naes, 1990) based on a technique developed by Ten Berge (Ten Berge, 1977). There the average scores y_{ij}^k are all multiplied by a constant c_{ik} which are optimized such that the scores for all products and attributes become as similar as possible. Brockhoff et al. (2015) present a mixed modelling based methodology to correct for the scaling effect. A mixed multiplicative model is introduced there and also a simpler alternative, the mixed assessor model (MAM), which we discuss in the following sections. According to Romano et al. (2008) both Ten Berge method and MAM outperform the method of standardized measurements. The Ten Berge approach is more restrictive than MAM in a way that there it is assumed that there is no disagreement effect. In MAM there are no such restrictions and the methodology of MAM opens up possibilities of being able to consider other effects such as carry-over, session/replication and others.

3.6.4 Mixed Assessor Model (MAM)

In standard 2-way mixed model considered in Equation 3.1 the interaction term d_{ij} is modeling the potential individual differences between the assessors in their scoring of the product differences. This includes as well differences in individual ranges of scale use (scale effect), as the real differences in perception of product differences (disagrement effect). Brockhoff et al. (2015) proposed a new mixed model approach, the Mixed Assessor Model (MAM), where the scaling effect is removed from the interaction term, so that the interaction term is modelling the real disagreements between assessors in scoring the products.

The Mixed Assessor Model (MAM) is given by

$$Y_{ijk} = \mu + a_i + \nu_j + \underbrace{\beta_i x_j}_{\text{scaling}} + \underbrace{d_{ij}}_{\text{disagreement}} + \varepsilon_{ijk} \tag{3.8}$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

where a_i is the assessor main effect, $i = 1, 2, ..., I$, the ν_j the product main effect, $j = 1, 2, ..., J$, $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and hence β_i is the individual (scaling) slope (the restriction $\sum_{i=1}^{I} \beta_i = 0$ is imposed in order to ensure that model 3.8 is uniquely parametrized). The d_{ij} term here captures interactions that are not scale differences hence "disagreements". In (Brockhoff et al., 2015) it was shown that MAM produces valid and improved hypothesis tests for as well overall product differences as post hoc product difference testing. Even though the detailed handling of the scaling part of model 3.8 is provided there together with the tool facilitating the analysis, the model and the tool have a number of strong limitations such as:

- can only handle balanced data
- can only consider one product effect
- can not handle other important effects such as, for example, replication/session effects, carry-over effects or others.

As pointed out in Section 3.5 and in paper in Appendix A it is sensible to consider more complex structures such as 3-way, where the replicate/session effect forms also part of the model as well as multi-way product structures

whenever it is possible to account for these effects. All that calls for a need in considering extended versions of MAM, where scaling effect can be part of a more complicated linear mixed effects model. It is also important to provide a tool that can handle the extensions together with handling unbalanced data.

3.6.5 Extended MAM

A function for performing inference for MAM as specified in Equation 3.8 is provided as a supplemental material for Brockhoff et al. (2015). However, as already mentioned, the function cannot handle unbalanced data. Even with one missing value, the tool can not be used even for a simple model as in Equation 3.8 and one should turn to a more general software, which is able to handle unbalanced data, fit MAM and provide inference tests. The **lme4** package Bates et al. (2013) is a nice open-source tool that can handle unbalanced data and fit LMM and therefore MAM as specified in Equation 3.8. As also pointed out in Brockhoff et al. (2015) one should use the sequential Type I tests in MAM because of the full confounding of the product ν_j and scaling $\beta_i x_j$ effects in the fixed part. The **lme4** package provides the Type I ANOVA table, however without the p values. The **lmerTest** package, presented in Section 5.1 and in paper in Appendix A generates the Type I ANOVA table together with the p values based on the Satterhtwaite's approximation to degrees of freedom. We hence use package **lmerTest** for the inference about the product and scaling effects for MAM as specified in Equation 3.8 as well as for the extended versions, which I specify in the following sections.

3.6.5.1 3-way MAM

In situations, where there are more complex replication structures (see Section 3.5) it is important to incorporate the replicate effect in the model, and if not significant, then subsequently eliminate it. The 3-way linear mixed assessor model, that incorporates the replicate effect as well, is defined in paper in AppendixC. For the `TVbo` data the 3-way model amounts to:

$$y_{ijk} = \mu + a_i + \nu_j + \underbrace{\beta_i x_j}_{\texttt{scaling}} + \underbrace{d_{ij}}_{\texttt{disagreement}} + r_k + ar_{ik} + a\nu_{jk} + \varepsilon_{ijk} \qquad (3.9)$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ r_k \sim N(0, \sigma^2_{replicate}),$$
$$ar_{ik} \sim N(0, \sigma^2_{assessor \times replicate}),\ \nu r_{jk} \sim N(0, \sigma^2_{product \times replicate}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

from which we may notice that three more random effects (replication effect and interactions between replication and the other effects) form part of the random structure compared to MAM in Equation 3.8. ν_j is again an effect corresponding to the product factor. In the unbalanced situations equation $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ for the product averages does not hold. Instead we calculate x_j as prediced values from a simple linear regression, where y is the response variable and ν is dependent variable. And then mean center the predicted values.

3.6.5.2 Multi-way product structure MAM

As already discussed in Section 3.5 frequently the products are formed of feautures. In the `TVbo` data, for example, the 12 products are formed as 4×3 combinations of Picture and TVset features. As stressed out in Section 3.5 and in paper on Appendix A it is important to account for such multi-way product structures, since additional important information on product similarities might be gained. In paper in Appendix C we propose the multifactorial product effect version of the MAM, where the product factor ν_j is replaced by the main feature effects and interactions between them. We show there, that this extended version provides more insight into the multi-way-product structure data and at the same time appropriately accounts for the scaling effect. In the `TVbo` data the multi-way product structure MAM, combined with the 3-way error structure is specified in the following form:

$$\begin{aligned} y_{ijkl} = \mu + a_i + \underbrace{\mathtt{TV}_j + \mathtt{Pic}_k + \mathtt{TVPic}_{jk}}_{\texttt{product effects}} + \underbrace{\beta_i x_{jk}}_{\texttt{scaling}} \\ + \underbrace{a\mathtt{TV}_{ij} + a\mathtt{Pic}_{ik} + a\mathtt{TVPic}_{ijk}}_{\texttt{disagreement}} \\ + r_l + ar_{il} + r\mathtt{TV}_{lj} + r\mathtt{Pic}_{lk} + \mathtt{TvPic}r_{jkl} + \varepsilon_{ijkl} \end{aligned} \tag{3.10}$$

$$
\begin{aligned}
a_i &\sim N(0, \sigma^2_{Assessor}) \\
a\texttt{Tv}_{ij} &\sim N(0, \sigma^2_{Assessor \times TV\,set}) \\
a\texttt{Pic}_{ik} &\sim N(0, \sigma^2_{Assessor \times Picture}) \\
a\texttt{TvPic}_{ijk} &\sim N(0, \sigma^2_{Assessor \times TV\,set \times Picture}) \\
r_l &\sim N(0, \sigma^2_{Repeat}) \\
ar_{il} &\sim N(0, \sigma^2_{Assessor \times Repeat}) \\
\texttt{Tv}r_{jl} &\sim N(0, \sigma^2_{TV\,set \times Repeat}) \\
\texttt{Pic}r_{kl} &\sim N(0, \sigma^2_{Picture \times Repeat}) \\
\texttt{TvPic}r_{jkl} &\sim N(0, \sigma^2_{TV\,set \times Picture \times Repeat}) \\
\epsilon_{ijkl} &\sim N(0, \sigma^2_{error})
\end{aligned}
\tag{3.11}
$$

Here x_{jk}, following the same arguments as for the 3-way MAM, are calculated as prediced values from a simple linear regression, where y is the response variable and `TV`, `Pic`, `TVPic` are the dependent variables. And then mean center the predicted values.

3.6.6 Automated analysis and MAM

Mixed assessor models in Equations 3.8, 3.9, 3.10 are still within the LMM class, therefore the methodology proposed in paper in Appendix A can still be applied here. Indeed from Equations 3.11 we observe that 9 random effects form the random part of the model in Equation 3.10. It might be that not all of these effects contribute to the systematic variation in the data and therefore could be excluded from the model. In the **SensMixed** package Kuznetsova et al. (2013b) the `step` method from the **lmerTest** package is used, that finds a parsimoneous random structure by sequentially removing non-significant random effects for extended MAMs considered in this chapter.

3.6.7 Post-hoc analysis for MAM

As discussed in Section 2.5, it is of particular importance to perform pairwise comparisons between the products. If the test for the product effect shows that the effect is significant, the post-hoc analysis can eventually detect which products do actually differ. There exist a large number of software tools that can perform pairwise comparisons: commercial as well as open-source. Some of them, like, for example, the open-source R package **lsmeans** Lenth and Hervé

(2014), are exclusively devoted to perform the post-hoc analysis. However, for MAM models, most of the software tools would not be able to perform pairwise comparison tests. The reason to that is connected with the fact that in MAM and the extensions product effect ν_j is fully confounded with the scaling effect $\beta_i x_j$, which can be easily seen from the definition of MAM and x_j in Equation 3.8. In paper in Appendix C the approach is presented for calculating the scale corrected pairwise comparisons between the products. The approach amounts to calculating the following test statistics:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{3.12}$$

where l is a contrast vector for the pairwise comparison in question. $\hat{\beta}$ are the fixed effects estimates. $\hat{C}$ is an estimated variance-covariance matrix of fixed effects. $\hat{\beta}$ are estimated from a model that does not contain the scaling effect. Note, that $\hat{C}$ here is estimated from a model with an additional scaling effect, that is from MAM model. By this, the contrast vector l is easily obtained and at the same time the scaling effect is corrected via $\hat{C}$.

Generally the t statistics does not follow the t distribution. As also discussed in Section 2.3, a method to find approximation to degrees of freedom can be used, such as Satterthwaite's and/or Kenward-Roger's (Giesbrecht and Burns, 1985; Kenward and Roger, 2009). In the **lmerTest** package the Satterthwaite's approximation is implemented (the algorithm is described Section 2.3.1). Here we use a modified version of it. Since the estimate of variance-covariance matrix $\hat{C} = C(\hat{\theta}_{\text{MAM}})$ is based on the estimated of variance-covariance parameters coming from MAM, that is $\hat{\theta}_{\text{MAM}}$, the calculation of $var(f(\theta_{\text{MAM}}))$ and A in Equation 2.10 are also based on estimates of variance-covariance parameters θ_{MAM} coming from MAM:

$$var(f(\theta_{\text{MAM}})) = l^\top C(\theta_{\text{MAM}})l \approx [\nabla_{f(\theta_{\text{MAM}})}\hat{\theta}_{\text{MAM}}]^\top A_{\theta_{\text{MAM}}}[\nabla_{f(\theta_{\text{MAM}})}\hat{\theta}_{\text{MAM}}] \tag{3.13}$$

By using the $var(f(\theta_{\text{MAM}}))$ and $A_{\theta_{\text{MAM}}}$ in the Satterthwaite's method, the denominator degrees of freedom are approximated and the corresponding p value for the test of product differences is obtained.

3.6.8 Examples

For illustrative purposes I consider here the `TVbo` data and the attribute `Contrast` as a response variable. This data are part of the **SensMixed** and **lmerTest** packages described in Chapter 5.

3.6.8.1 3-way MAM

First, I consider a 3-way model as specified in Equation 3.9, but without the scaling term $\beta_i x_j$.

Table 3.1: Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom for 3-way model and Contrast attribute

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
product	105.41	9.58	11	77.00	8.40	<0.001

From the `anova` output it can be seen that the `product` effect is significant. Then, I add the scaling effect into the fixed part which results in construction of model in Equation 3.9, and again apply the `anova` method, but specify that it should produce the sequential ANOVA table (Type I):

Table 3.2: Analysis of Variance Table of type I with Satterthwaite approximation for degrees of freedom for MAM in Equation 3.9 and Contrast attribute

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
product	143.96	13.09	11	70.00	11.47	<0.001
Assessor:x	40.13	5.73	7	70.00	5.02	<0.001

From the output in Table 3.2 it can be seen that the `product` effect is significant. Actually the p value for the product effect became lower than in the previous model without the scaling effect, but it is still within the same order of magnitude. It can be also seen that the scaling effect, that is the `Assessor:x` term, is significant.

3.6.8.2 multi-way product structure MAM

The `TVbo` data has a multi-way product structure, where produts are 3-by-4 combinations of TVset and Picture feauteres. Hence, as also discussed in paper in Appendix C, it can be sensible to consider two main effects corresponding to TVset and Picture features and interaction between them. Model, that has a multi-way product structure in addition to the 3-way error structure is constructed and `anova` from the **lmerTest** package is applied then:

Table 3.3: Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom for 3-way multi-way product structure TVbo data for Contrast attribute

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
TVset	28.66	14.33	2	14.00	12.56	< 0.001
Picture	11.66	3.89	3	63.00	3.41	0.02290
TVset:Picture	26.43	4.41	6	63.00	3.86	0.00242

From the `anova` output in Table 3.3 it can be observed that the main effects as well as the interaction effect are significant. Then I add the scaling effect which results in construction of the MAM as specified in Equation 3.10. Then the sequential ANOVA table is generated via **lmerTest**:

Table 3.4: Analysis of Variance Table of type I with Satterthwaite approximation for degrees of freedom for MAM in Equation 3.10 and Contrast attribute

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
TVset	25.09	12.55	2	6.30	11.00	0.00881
Picture	13.25	4.42	3	50.54	3.87	0.01441
TVset:Picture	30.04	5.01	6	50.54	4.39	0.00122
Assessor:x	15.83	2.26	7	43.78	1.98	0.07950

From the output in Table 3.4 it can be seen that the p value for the interaction and for the `Picture` effects become slightly smaller compared to the previous model without the scaling effect. the p value for the `TVset` effect becomes slightly bigger than for the one without the scaling effect. The p value for the scaling effect is now slightly bigger than in a one-way product model, but still the effect should be kept according to Brockhoff et al. (2015), where the authors suggest to keep the scaling effect whenever its p value is less than 0.2.

3.6.8.3 Automated MAM

Here, followong the methodology presented in paper in Appendix A, I consider the full model as specified in Equation 3.10. Then the the simplification of the random structure of the MAM is performed in Table 3.5 using the `step` function from the **lmerTest**. Finally the sequential ANOVA is obtained in Table 3.6.

Table 3.5: Likelihood ratio tests for the random-effects and their order of elimination representing Step 1 of the automated analysis for the TVbo data for attribute Contrast

	χ^2	Chi.DF	elim.num	p-value
Assessor:Picture	0.00	1	1	1.00000
Repeat:Picture	0.00	1	2	1.00000
Repeat:TVset	0.00	1	3	1.00000
Repeat:TVset:Picture	0.00	1	4	1.00000
Assessor	0.00	1	5	0.97656
Repeat	0.06	1	6	0.80826
Assessor:TVset	8.19	1	kept	0.00421
Assessor:Picture:TVset	2.72	1	kept	0.09895
Assessor:Repeat	26.27	1	kept	< 0.001

Table 3.6: Analysis of Variance Table of type I with Satterthwaite approximation for degrees of freedom for reduced MAM for Contrast attribute

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
TVset	27.71	13.85	2	15.51	8.31	0.00352
Picture	19.84	6.61	3	149.06	3.97	0.00935
TVset:Picture	44.98	7.50	6	149.06	4.50	< 0.001
Assessor:x	23.74	3.39	7	117.99	2.03	0.05624

First we observe, that 6 random effects were eliminated from the initial model as being non-significant according to the 0.1 Type 1 error rate (the default one in the `step` function in **lmerTest** in tests for the random effects). From ANOVA table presented in Table 3.6 it can be seen that the p values for both scaling as well as the product effects became smaller compared to the ones from Table 3.4.

Chapter 4

Visualizing results in sensory and consumer studies

It is of particular importance to develop tools for simple and visual interpretation of the results based on the statistical methods developed. In particular this is important for ensuring efficient communication between the statistician and the user of the methods.

The ConsumerCheck software tool presented in paper in Appendix C provides a number of visualisation tools dedicated for analysis of consumer liking data. The plots for exploratory analysis are: *Box plots* for the liking scores, *stacked histogram*, where the distribution of scores of consumers is visualised in a more detailed way than in *Box plots*. The multivariate-based plots are *PCA* and *Preference mapping*. The *Preference mapping* is a valuable tool that links together sensory characteristics of the products together with the consumer likings. Pairwise-comparisons plots for tests of product differences that are based on mixed model ANOVA form also part of the software - we present them in the following sections.

The widely used open source software PanelCheck Nofima Mat (2008) contains as well a number of univariate as well as multivariate tools for detecting and vizual-

ising individual differences between assessors in sensory studies. The multivariate tools povided there comprise *Tucker-1* and *Manhattan plots*. *Tucker-1* plot provides information about reproducibility across assessors as well as systematic variation for each assessor. *Manhattan plot* is a screening tool, that provides a quick information data structure for each assessor. Univariate ANOVA-based plots are the so-called *F plot*, *MSE plot* and *p∗ MSE plot*. Other so-called "simple methods", provided in PanelCheck are *Eggshell plot*, *Profile plot*, *Correlation plot*. The *Correlation* and the *Profile plots* show how each assessor uses scale as well as how each assessor percept products. The "simple methods" are a good choice for exploratory analysis of sensory data, which is very important step in data analysis. The advantage of ANOVA based tools is that they produce hypothesis tests as well as confidence intervals for product differences.

The focus of this chapter is on presenting mixed modelling ANOVA-based visualisation tools developed in this project and presented in papers in Appendices A and C. For illustration, I will in this chapter use the `TVbo` data presented in Section 3.1.1.1, which has a multi-way product structure.

It is crucial to provide an intuitive and user-friendly tool, that can facilitate analysis as well as representation of the results. ConsumerCheck software contains a graphical user interface (GUI) and indeed is a nice and easy-to-use tool for non-statisticians. The **SensMixed** contains an application, that has also GUI and such functionalities as importing data, saving results in different formats and others. Hence, visualisation tools that I present in the following section can be easily used by non-statisticians/sensory practitioners.

4.1 Multi-attribute plots in sensory data

4.1.1 Multi-attribute plot for the product effect as in PanelCheck

Since the `TVbo` data is balanced, the PanelCheck software can be used. The inbuilt mixed modelling ANOVA results are visualized there by multi-attribute bar plots of F-statistics combined with colour coding of the significance results. In this way the F-statistic is used as a kind of effect size measure. The following 3-way model was constructed for each attribute:

$$y_{ijk} = \mu + a_i + \nu_j + d_{ij} + r_k + ar_{ik} + a\nu_{jk} + \varepsilon_{ijk} \tag{4.1}$$

where a corresponds to the assessor effect, d corresponds to the interaction between assessor and product effects and r to the replicate effect.

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{product \times assessor}),\ r_k \sim N(0, \sigma^2_{replicate}),$$
$$ar_{ik} \sim N(0, \sigma^2_{assessor \times replicate}),\ \nu r_{jk} \sim N(0, \sigma^2_{product \times replicate}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

The F statistics in test for the product effect is then:

$$F_{product} = \frac{MS_{product}}{MS_{product \times assessor} + MS_{product \times replicate} - MSE} \tag{4.2}$$

The following plot represents the bars for the square root of the F statistics in test for the product effect based on models from Equation 4.1. The colour of the bars represents the significance levels. This plot is valuable in detecting whether assessors are able to discriminate between the products according to an attribute in question. If the tests show that the assessors are not able to discriminate the products for some attribute in question, then it is a common thing to remove the attribute from the studies, as the assessors seem not to being able to discriminate between the products.

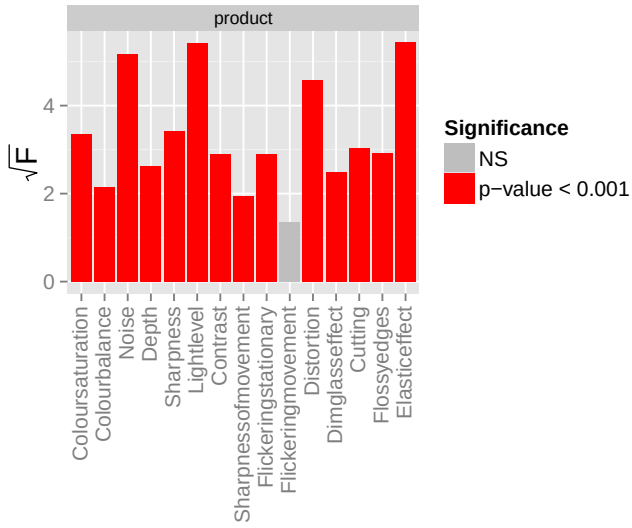


Figure 4.1: structure of a typical sensory data set

From the plot it can be observed that for almost all the attributes the product effect is highly significant. Only for the `Flickeringmovement` attribute the plot claims that the product effect is non-significant according to the 0.05 Type 1

error rate. Therefore, according to this plot, attribute `Flickeringmovement` can be excluded from further analysis, since the assessors seem to be not able to discriminate the products according to it. For the convenience purposes, the plot was made using the **SensMixed** package, however the PanelCheck software produces a similar one.

4.1.2 Multi-attribute plot for the product effect in SensMixed

As it is pointed out in Brockhoff et al. (2015) it is important to account for the the scaling effect whenever it is possible, since then the tests for the product effects become more powerful. In **SensMixed** application the multi-attribute plots for tests of product effects based on mixed assessor models (MAM) and extensions (see Section 3.6 and paper in Appendix C) can be easily generated. For the example considered in this chapter, the models are constructed as specified in Equation 3.9, which include the scaling effect compared to the model in Equation 4.1. The multi-attribute plot for the product effect is then the following one:

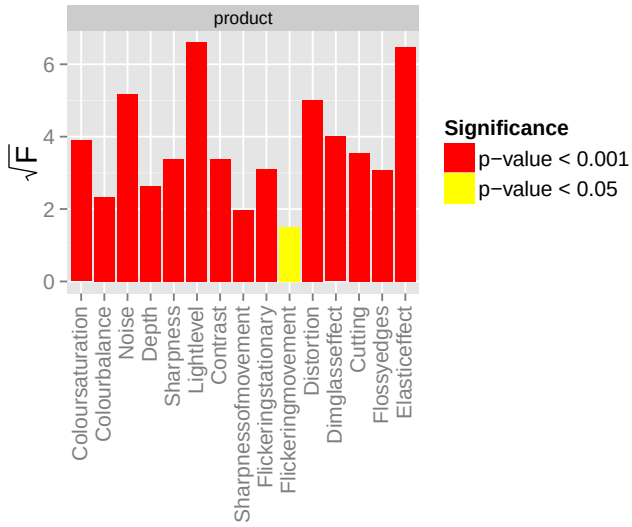


Figure 4.2: structure of a typical sensory data set

From the plot it is clear that `Flickeringmovenent` attribute is significant. This is because the $MS_{product \times assessor}$ in denominator in Equation 4.2 became smaller and now stands for the real disagreement between assessors, hence the $F_{product}$ statistics became bigger. According to this new plot, that the **SensMixed** application provides, we should not exclude this attribute from further analysis.

4.1.3 $\tilde{\delta}$ plot in SensMixed

The F *plots* can be a good approach, especially within PanelCheck, where the multi-attribute bar plot of the overall product differences are used only for single-factor product effects and with the same choice of F-test denominator across all the attributes of a plot. The `TVbo` data has a multi-way product structure: the products are essentially 3-by-4 combinations of `TVset` and `Picture` features. As it is stressed in paper in Appendix A, multi-way product structure should be taken into account, since it can provide additional insight into the data. Moreover, with the automated approach proposed in the paper in Appendix A different effects may have different noise structures, that is, different factors may be tested using different F-test denominators, so that heights of the bars corresponding to Fs or $\sqrt{F}$s can not be directly compared.

In paper in Appendix E the so-called $\tilde{\delta}$-primes effect sizes are proposed, which are the average of all standartized pairwise differences between products. For a one product factor situation with I levels in product factor the $\tilde{\delta}$-primes are defined there in the following way:

$$\tilde{\delta} = \sqrt{\frac{2}{I(I-1)} \sum_{i_1 < i_2}^{I} (\frac{\mu_{i_1} - \mu_{i_2}}{\sigma})^2} \tag{4.3}$$

where $\sum_{i_1<i_2}^{I}$ means the sum of all unique combinations of the two indeces, μ_{i_1} is a mean for product i_1 .

For the 2-way product structure case with A and B factors corresponding to two features of the products the $\tilde{\delta}$ are specified as follows:

$$\tilde{\delta}_A = \sqrt{\frac{2}{I(I-1)} \sum_{i_1 < i_2}^{I} (\frac{\alpha_{i_1} - \alpha_{i_2}}{\sigma})^2} \tag{4.4}$$

$$\tilde{\delta}_B = \sqrt{\frac{2}{J(J-1)} \sum_{j_1<j_2}^{J} (\frac{\beta_{j_1} - \beta_{j_2}}{\sigma})^2} \tag{4.5}$$

$$\tilde{\delta}_{A\times B} = \sqrt{\frac{2}{IJ(IJ-1)} \sum_{ij_1<ij_2} (\frac{\gamma_{ij_1} - \gamma_{ij_2}}{\sigma})^2} \tag{4.6}$$

where α is a main effect for factor A, β is a main effect for factor B and γ is an interaction effect. Note, that the interaction effect γ represents the pure interaction, that is, in the 2-way case they are:

$$\gamma_{ij} = \overline{y}_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}_{\cdot\cdot}$$

so the main effects are cancel out from the interaction effect. In that way $\tilde{\delta}$ can be compared among all possible effects.

It is shown there that the effect size estimate $\tilde{\delta}$ can be interpreted and compared across any attributes and factor levels especially when there is a multi-way product structure, with different number of levels and different number of observations within the levels.

The unbiased estimate of the population $\tilde{\delta}$-value is provided in the paper, however it can be obtained only for the balanced situations. One of the main features of the **SensMixed** package is that it supports unbalanced data. For the purposes of providing a common framework for getting the estimates of the $\tilde{\delta}$ for both balanced and unbalanced data, the estimates there are calculated as the standartized average differences of the least squares means. So, basically, substituting parameters in Equation 4.3 by their estimates coming from a mixed effects model in question. For instance, for the `TVbo` data, by substituting αs, βs and γs in Equations 4.4, 4.5 and 4.6, with the corresponding least squares means, σ with its estimate $\hat{\sigma}$ we get:

$$\hat{\tilde{\delta}}_{TV\,set} = \sqrt{\frac{1}{3} \sum_{i_1<i_2}^{I} (\frac{\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2}}{\hat{\sigma}})^2} \tag{4.7}$$

$$\hat{\tilde{\delta}}_{Picture} = \sqrt{\frac{1}{6} \sum_{j_1<j_2}^{J} (\frac{\hat{\beta}_{j_1} - \hat{\beta}_{j_2}}{\hat{\sigma}})^2} \tag{4.8}$$

$$\hat{\tilde{\delta}}_{TVset \times Picture} = \sqrt{\frac{1}{66} \sum_{ij_1 < ij_2} (\frac{\hat{\gamma}_{ij_1} - \hat{\gamma}_{ij_2}}{\hat{\sigma}})^2} \tag{4.9}$$

The following plot represents the $\tilde{\delta}$ plot generated by **SensMixed**. The first advantage of this plot compared to the one-way product plot in Figure 4.1 is that it provides information also on the features `Picture` and `TVset`. For example, for `Colourbalance` attribute it is mainly due to the `Picture` feature the products differ between each other. Since the $\tilde{\delta}$ primes represent the effect sizes, the sizes of the bars can be compared between each other. So, for instance, the size of the `Picture` effect for the `Lightlevel` attribute is much higher than for the other attributes.

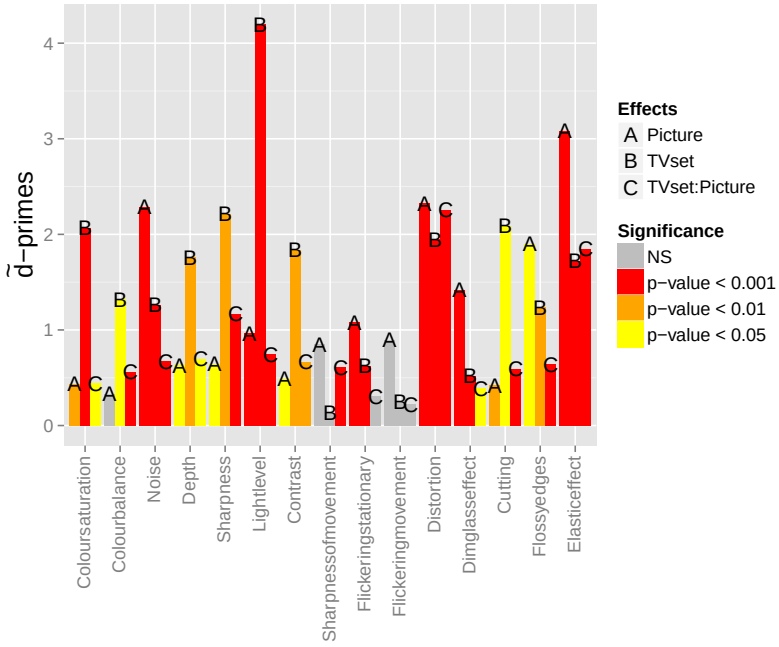


Figure 4.3: Bar-plot for the $\hat{\tilde{\delta}}$s for the product effects. TVbo data

4.1.4 Multi attribute plot for the scaling effect in SensMixed

The plot for the scaling effect represents the square root of the F statistics in test for the fixed scaling effect. The plot is valuable in providing an overall information regarding whether there is a scaling effect for an attribute in question. From the following plot we may see that for most of the attributes the p values for the Scaling effect are less than 0.05. For a number of effects the the p values are higher than 0.05, however, according to Brockhoff et al. (2015) the scaling effect should be accounted in the model whenever its p value is less than 0.2.

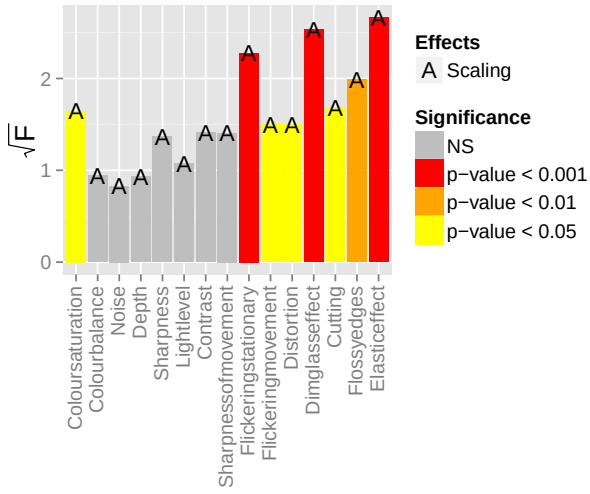


Figure 4.4: Bar-plot for the $\sqrt{F}$ statistics in test for the scaling effect. TVbo data

4.1.5 Multi-attribute plot for the random effects

$\sqrt{\chi^2}$ plot represents the bars for the square root of the χ^2 statistics of the likelihood ratio test applied to random-effects for each sensory attribute. The colours of the bars represent the significance level of the effects. This plot is a valuable visualisation tool that helps the user to quickly investigate, for instance, whether there is a replication effect, or is there a disagreement between assessors on scoring the products and if yes, then according to which features. If there is a requirement for the reduction of the random effects, then the $\sqrt{\chi^2}$values

are the sequential ones, that is they come from the stepwise selection process based on the methodology proposed by paper in Appendix A. The following plot represents the sequential $\sqrt{\chi^2}$ values. From the plot we may, for instance, observe that Assessors disagree in their scoring mainly due to the `Picture` feature.

We may notice that the interaction between `Repeat` and product effects are not present here (they were not included in the analysis since the one-way product analysis showed that they are not significant).

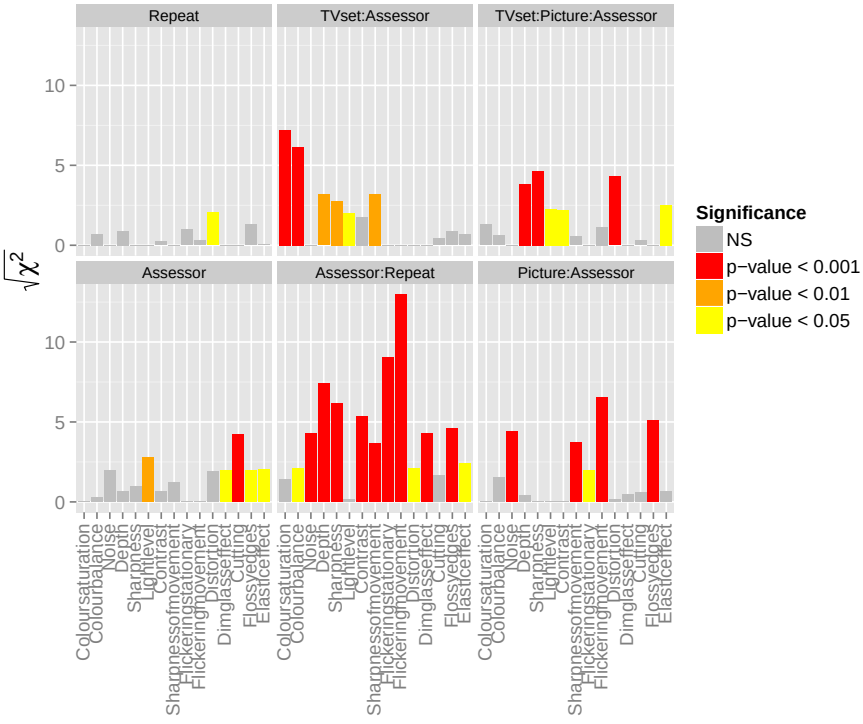


Figure 4.5: Barplots for $\sqrt{\chi^2}$-statistics of likelihood ratio test for random-effects for the TVbo data

4.2 Post-hoc plots

As stressed in Section 2.5 it is of particular importance to perform post-hoc testing in ANOVA modelling. Indeed, the F plots provide information on the overall product difference. It is of interest to know more about which products are different from each other. Are all of them different or is it just a clear difference between two of the products? Both ConsumerCheck as well as SensMixed provide the pairwise comparisons together with confidence intervals. For example, in the `TVbo` data the products differ according to the **TVset** feature for attribute `COlourbalance`. The following plot shows the pairwise compaisons between the main effect corresponding to `TVset` feature. We can dee that there is a significance difference between products `TVset 1 - TVset 2` as well as `TVset 2 - TVset 3`. In ConsumerCheck the p values are Bonferroni corrected (Dunn, 1961). We leave for the future work to implement a correction in **SensMixed**.

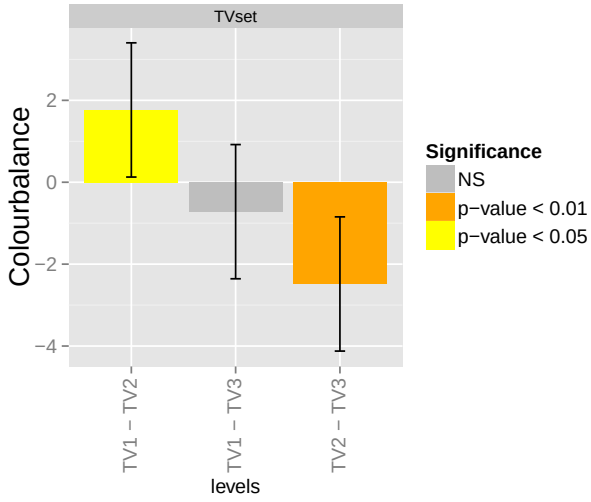


Figure 4.6: Pairwise comparisons between levels of TVset main effect. TVbo data

Chapter 5

Software tools

5.1 R package lmerTest

There are a number of statistical software packages containing routines for LMMs. These include, for instance, SAS, SPSS, STATA, R and others. The major advantage of R R Core Team (2015) is that it is a freely available, dynamically developing, open-source environment for statistical computing and graphics. There are few R packages that can be used to fit LMMs. The most famous ones are the **nlme** Pinheiro et al. (2015) and **lme4** Bates et al. (2013). The **lme4** package uses efficient linear algebra as implemented in the **Eigen** package, which is a C++ template library for linear algebra. As a consequence, the **lme4** is faster than **nlme**. Another advantahge of **lme4**, is that it can easily handle multiple crossed random effects and provides a very user-friendly interface for specifying LMMs. The downside **lme4** compared to **nlme** is that it cannot model heterocedascity nor correlation in residuals (at least currently). However, most of the covariance structures, that are commonly used in sensory and consumer studies (see Chapter 3), the **lme4** package can easily handle.

One of the main focuses of this project was to provide user-friendly tools for testing and model building LMMs dedicating for the conjoint analysis as part of the ConsumerCheck software (see paper in Appendix D). It was decided that the core of the ConsumerCheck software should be written using Python Foundation

(Foundation) language and the conjoint analysis functionailties of the software should be written in R, since it is an open-source software, specifically dedicated for statistical analysis and the functions coming from R can be easily called from Python via library **pypeR** .

At the time when the development of the conjoint analysis functionalities started (back in September 2010), it was decided to base the conjoint analysis on the **lme4** package, that is the LMMs should be fitted via **lme4** package. A number of things were not available at that time: p values with approximation methods for the tests of fixed effects, stepwise building approaches, post-hoc analysis and others. Some of them can be found in nowadays via different packages. For example, the Lenth and Hervé (2014) package can perform the post-hoc analysis, the **pbkrtest** Halekoh and Højsgaard (2014b) can perform tests on the fixed effects e t.c.

The main functions that the **lmerTest** package consists of together with the examples illustrating them are introduced in paper in Appendix B. The focus of this chapter is on presenting the structure of the **lmerTest** package, introduction to the functions and some implementational details on the functions.

5.1.1 Outline of the lme4 package

The **lme4** package is a well-known and widely used R-package designed to fit linear as well as non-linear mixed effects models. Some of the **lme4** package main strengths are the user-friendly interface, the ability to handle unbalanced data, multiple crossed effects and being very fast even for large data sets. 115 packages are built on top of the **lme4** package and contain different functionalities.

The **lmerTest** is one of those packages that are built on top of the **lme4** package and includes a number of functions that facilitate model building, tests for the fixed as well as random effects for objects of class `lmerMod`, which fits linear mixed effects.

5.1.1.1 "lmer"-model formulas

All the models are constructed using the same principle. As an example let `response` be the response variable, `eff1`, `eff2` be the main fixed-effects, their interaction would then be `eff1:eff2`, and `eff3` - the random-effect. An `lmer` model to analyze them would then be:

```
modelEx <- lmer(resp ~ eff1*eff2 + (1 | eff3))
```

It can be seen that the model formula consists of two expressions separated by the ∼ symbol in the left hand side of the formula the response variable `resp` is specified. The right-hand side consists of one or more terms separated by `'+'` symbols. The first term is `eff1*eff2`, which represents three fixed effects: two main effects `eff1`, `eff2` and interaction between them `eff1:eff2` (symbol `'*'` means all main effects plus all possible interactions between them). `(1 | eff3)` is a specification of a random-effect. In general specification of each random-effects term consists of two expressions separated by the vertical bar, `'|'`, symbol and enclosed in parentheses. The expression on the right of the `'|'` is a factor (here `eff3`). In a scalar random-effects term (which we have in our case), the expression on the left of the `'|'` is '1'. Such a term generates one random-effect (i.e. a scalar) for each level of the factor. Another possibilities for the expression on the left of the `'|'` could be `1 + eff4`, which constructs random coefficient model with correlation between slope and intercept.

5.1.2 Structure of the lmerTest package

In R all objects belong to some class. For instance, a number belongs to the class "numeric", a string belongs to the class "character". In the **lme4** the main class is `merMod`, which contains another two classes: `lmerMod` and `glmerMod` package objects. `lmerMod` stands for linear mixed effects models, whereas `glmerMod` stands for generalized linear models. Objects, specified via function `lmer` (see Section 5.1.1.1 for the definition) belong to the `lmerMod` class. In the **lmerTest** package the class `merModLmerTest` is defined, that contains the `lmerMod` class via the following code:

```
merModLmerTest <- setClass("merModLmerTest",
                           contains = c("merMod", "lmerMod"))
```

This means that whenever the package **lmerTest** is attached (is "used"), all objects created via `lmer` function belong to the class `merModLmerTest`. I can verify that in the following example. I consider the sensory data with the name `TVbo`, which is introduced in Section 3.1.1.1. For illustration, I choose the attribute `Coloursaturation` as a response variable, in the fixed effects I consider two main effects corresponding to features `TVset` and `Picture` and an effect, corresponding to an interaction between them. I consider `Assessor` as random effect. For the discussions regarding the specifications of LMMs for sensory data I refer to Sections 3.2 and 3.5. In the following code I attach the **lme4** package

and specify the model via `lmer` function, as exlained in Section 5.1.1.1 and then call `class` function, which gives the class, that the LMM object belongs to:

```
library(lme4)
m.lmerMod <- lmer(Coloursaturation ~ TVset*Picture +
                  (1|Assessor), data = TVbo)
class(m.lmerMod)


## [1] "lmerMod"
## attr(,"package")
## [1] "lme4"
```

It can be seen, that the class of the `m.lmerMod` object is `lmerMod`. In the following I attach the **lmerTest** package and fit the same LMM model via `lmer` function and then check the class of the fitted object:

```
library(lmerTest)
m.merModLmerTest <- lmer(Coloursaturation ~ TVset*Picture +
                  (1|Assessor), data = TVbo)
class(m.merModLmerTest)


## [1] "merModLmerTest"
## attr(,"package")
## [1] "lmerTest"
```

The class of the fitted object `m.merModLmerTest` is `merModLmerTest`.

Specifying the `merModLmerTest` class in such way gives few privileges, namely, all functions, defined in the **lme4** package for extracting information on an `lmer` objects can be:

- directly used via `merModLmerTest`
- respecified via `merModLmerTest`

In Figure 5.1 the structure of the package in relation to the **lme4** package can be viewed. For instance, the arrow from `lmerMod` to `fixef` means that the function with the name `fixef` can be applied on objects of class `lmerMod` - this function extracts estimates of the fixed effects. It can be seen that an arrow

goes from `merModLmerTest` to `lmerMod`: this means that all functions which can be applied on objects of class `lmerMod` can be also applied on objects of class `merModLmerTest`. Arrows, coming from the `merModLmerTest` to `anova` and `summary`, mean these functions are respecified. For functions `step`, `rand` and `lsmeans` arrows go only from the `merModLmerTest` class. This means these functions are only defined for the `merModLmerTest` class.

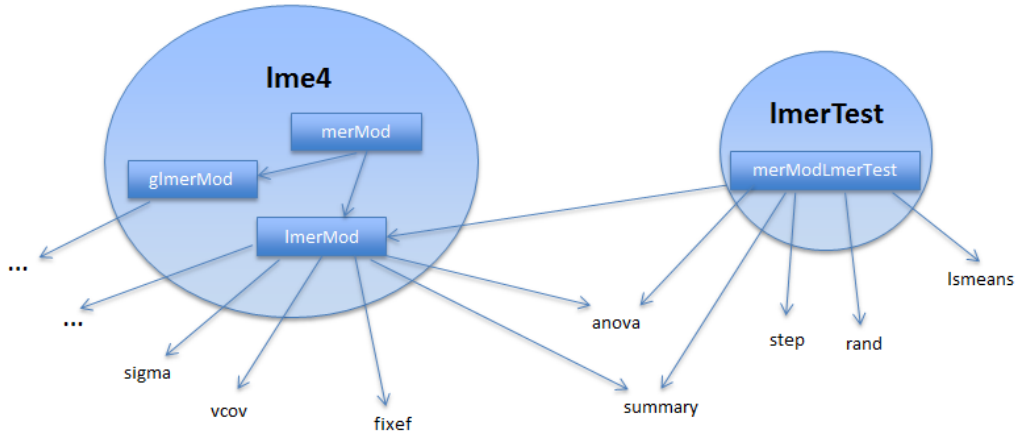


Figure 5.1: Structure of the lmerTest package

The reason why we have decided to create the `merModLmerTest` class is related to the fact, that we wanted to maintain the user-friendliness of the **lme4** package. The **lme4** package is indeed widely used, and the `summary` and the `anova` functions are the two main functions providing inference on the parameters of an LMM model. Therefore, from a user perspective, it is crucial to being able to get tools for testing the effects through the same functions, namely `anova` and `summary` functions.

5.1.3 summary and lsmeans/difflsmeans functions

The `summary` function applied to an object of class `lmerMod` prints the summary on the estimates of the parameters of the fitted LMM model. The information on the estimates of the fixed effects, as part of the output of the `summary` function can be extracted via the `coefficients` function in the following way:

```
coefficients(summary(m.lmerMod))
```

```
##                    Estimate Std. Error t value
## (Intercept)           7.144      0.340  20.998
## TVsetTV2              2.681      0.405   6.625
## TVsetTV3              0.413      0.405   1.019
## Picture2              0.169      0.405   0.417
## Picture3              0.313      0.405   0.772
## Picture4              1.044      0.405   2.579
## TVsetTV2:Picture2    -0.325      0.572  -0.568
## TVsetTV3:Picture2    -0.344      0.572  -0.601
## TVsetTV2:Picture3    -0.988      0.572  -1.725
## TVsetTV3:Picture3    -0.269      0.572  -0.470
## TVsetTV2:Picture4    -0.400      0.572  -0.699
## TVsetTV3:Picture4    -1.237      0.572  -2.162
```

From the output it is seen that the estimates of fixed effects are provided together with their standard deviation and the t statistics. However, there are no p values in the output, so it is hard to judge about the significance of the effects.

Now I apply the same functions to the `m.merModLmerTest` object:

```
coefficients(summary(m.merModLmerTest))
```

```
##                    Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)           7.144      0.340  49.8  20.998 0.00e+00
## TVsetTV2              2.681      0.405 173.0   6.625 4.22e-10
## TVsetTV3              0.413      0.405 173.0   1.019 3.09e-01
## Picture2              0.169      0.405 173.0   0.417 6.77e-01
## Picture3              0.313      0.405 173.0   0.772 4.41e-01
## Picture4              1.044      0.405 173.0   2.579 1.07e-02
## TVsetTV2:Picture2    -0.325      0.572 173.0  -0.568 5.71e-01
## TVsetTV3:Picture2    -0.344      0.572 173.0  -0.601 5.49e-01
## TVsetTV2:Picture3    -0.988      0.572 173.0  -1.725 8.62e-02
## TVsetTV3:Picture3    -0.269      0.572 173.0  -0.470 6.39e-01
## TVsetTV2:Picture4    -0.400      0.572 173.0  -0.699 4.86e-01
## TVsetTV3:Picture4    -1.237      0.572 173.0  -2.162 3.20e-02
```

As can be observed, the output provides two more columns: "df" standing for the degrees of freedom and $Pr(> t)$ standing for the p value from the t test,

where the "df" were used as the degrees of freedom. In the **lmerTest** package the t value for the `summary` function is calculated using the Equation 3.12, where l is an identity vector. For instance, of an intercept in this example the vector is:

$$l = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

The "df" are found from the method, known as Sattherthwaite's method of approximation. The algorithm of the method as applied to the t test is described in Section 2.3.1 and some details on the implementation of the method in the **lmerTest** package are given in the following sections. In this examples, one generally would be more interested in comparing the levels of the factors (`TVset` and `Picture`), or calculating the mean scores for each level of the factors. One may, of cource, calculate the means and then compare them, but it is of interest to also perform tests. In the **lmerTest** package two functions are provided: `lsmeans` and `difflsmeans`. `lsmeans` calculate the so-called least squares means, presented by Harvey (1975), with confidence intervals whereas `difflsmeans` calculate the differences between the least squares means together with the confidence intervals. The tests for the least squares means actually amount to again performing the t test, but with the different l vector. In the **lmerTest** the l vectors for the `lsmeans` function use the `popMatrix` function from the **doBy** package (Højsgaard et al., 2014b). The l vectors for differences of least square means are then constructed as pairwise differences of ls vectors from the least square means. The confidence intervals are calculated based on Equation 2.14. In this example the following code will calculate the least squares means for an effect `TVset` together with the confidence intervals:

```
lsmeans(m.merModLmerTest, test.effs = "TVset")


## Least Squares Means table:
##            TVset Picture Estimate Standard Error   DF t-value Lower CI
## TVset  TV1   1.0      NA    7.525          0.233 12.4    32.3     7.02
## TVset  TV2   2.0      NA    9.778          0.233 12.4    42.0     9.27
## TVset  TV3   3.0      NA    7.475          0.233 12.4    32.1     6.97
##            Upper CI p-value
## TVset  TV1     8.03  <2e-16 ***
## TVset  TV2    10.28  <2e-16 ***
## TVset  TV3     7.98  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The least squares means in the balanced situation are equal to the corresponding means.

5.1.4 anova function

`anova` function applied to `lmer` object of class `lmerMod` (from the **lme4** package produce the sequential ANOVA table. For example, if I apply `anova` on the `m.lmerMod` object I get the following output:

```
anova(m.lmerMod)


## Analysis of Variance Table
##               Df Sum Sq Mean Sq F value
## TVset          2  221.5   110.8   84.53
## Picture        3   11.2     3.7    2.85
## TVset:Picture  6   13.8     2.3    1.76
```

The output shows `Df` refering to numerator degrees of freedom of the corresponding F statistics (`F value` in the fourth column). `Sum Sq` refers to the sequential sums of squares (see Section 2.4 for the definition). `Mean Sq` refers to the mean squares, which are calculated in the following way: $\texttt{Mean Sq} = \frac{\texttt{Sum Sq}}{\texttt{Df}}$. Finally, the `F value` corresponds to the F statistics. Due to the reasons discussed in Section 2.3, the F statistics generally follows an unknown distribution. This is one of the reasons why the authors of the **lme4** package do not provide the p values. The only way to judge about the effect sizes from this table is by looking at the magnitude of `F value` and `Mean Sq`. From this output it seems like the effect of `TVset` feature is high, whereas the effect sizes of `Picture` and `Picture:TVset` seem to be small.

In the **lmerTest** the Satterthwaite's approximation method via algorithm proposed by Fai and Cornelius (1996) and presented in Section 2.3.1. There it is assumed, that the F value follows the F distribution and the denominator degrees of freedom are approximated. The implementation of the Satterthwaite's method is also wrapped to `anova` function. In this example, applying `anova` function to the object `m.merModLmerTest` results in the generation of the following ANOVA table:

```
anova(m.merModLmerTest)


## Analysis of Variance Table of type III  with  Satterthwaite
## approximation for degrees of freedom
##               Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)
## TVset          221.5   110.8     2   173    84.5 <2e-16 ***
## Picture         11.2     3.7     3   173     2.9  0.039 *
```

```
## TVset:Picture   13.8     2.3     6   173     1.8  0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen that the output now is slightly extended. The additional column has a name `DenDF` and produces denominator degrees of freedom calculated based on Satterthwaite's method. `Pr(>F)` corresponds to the p value, calculated based on the F test with the `DenDF` used as the denominator degrees of freedom. The resulting p values indicate that the `TVset` effect is highly significant. The `Picture` effect is also significant, but not the interaction effect.

The heading of the output also has changed. Now it states : Type III ANOVA table. This means that the Type III hypothesis tests, as described in Section 2.4 were performed here. Following Table 2.6 the Type III hypothesis test for the main effect `TVset` in this example is:

$$\alpha_i - \alpha_{i'} + (1/4)\sum_j (\alpha\beta_{ij} - \alpha\beta_{i'j}) = 0 \qquad \forall i, i' \tag{5.1}$$

The test for the main effect `Picture` is:

$$\beta_j - \beta_{j'} + (1/3)\sum_i (\alpha\beta_{ij} - \alpha\beta_{ij'}) = 0 \qquad \forall j, j' \tag{5.2}$$

The test for the `TVset:Picture` effect is:

$$\alpha\beta_{ij} - \alpha\beta_{i'j'} = 0 \qquad \forall i, i', j, j' \tag{5.3}$$

All software add constraints to parameters in order to guarantee a unique solution. The default in R is treatment contrasts, where the first level of any factor is set to zero. Under the treatment contrats, the test for the difference between `TVset 2` and `TVset 1` reduces to:

$$\alpha_2 + (1/4)(\alpha\beta_{22} + \alpha\beta_{23} + \alpha\beta_{24}) = 0$$

And the test for the difference between `TVset 3` and `TVset 1` reduces to:

$$\alpha_3 + (1/4)(\alpha\beta_{32} + \alpha\beta_{33} + \alpha\beta_{34}) = 0$$

Then the Type III hypothesis contrasts matrix for the main `TVset` effect under the treatment contrasts amounts to:

$$L = \begin{array}{l} \\ \texttt{TVset 2 - TVset 1} \\ \texttt{TVset 3 - TVset 1} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} \mu & \alpha_2 & \alpha_3 & \beta_2 & \beta_3 & \beta_4 & \alpha\beta_{22} & \alpha\beta_{32} & \alpha\beta_{23} & \alpha\beta_{33} & \alpha\beta_{24} & \alpha\beta_{34} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0 & 0.25 \end{pmatrix} \end{array}$$

We can verify that this contrast matrix produces the same F statistic as the `anova` method above:

```
L <- matrix(c(0, 1, 0, 0, 0 ,0, 0.25, 0, 0.25, 0, 0.25, 0,
              0, 0, 1, 0, 0, 0, 0, 0.25, 0, 0.25,0, 0.25),
            nrow=2, ncol=length(fixef(m.lmerMod)), byrow = TRUE)
C <- as.matrix(L %*% vcov(m.lmerMod) %*% t(L))
q <- qr(C)$rank
F.stat <- (t(L %*% fixef(m.lmerMod)) %*%
            solve(C) %*% (L %*% fixef(m.lmerMod)))/q
F.stat


##      [,1]
## [1,] 84.5
```

The construction of the L contrasts matrix depends on the constraints that are put on model paramaters. In our implementation the calculation of the ANOVA table is invariant with respect to the constraints. This is simply achived by refitting the model with the `contr.SAS` contrasts, which corresponds to setting the last parameter to 0. If one uses `Anova` function from the **car** package Fox and Weisberg (2011), one needs to use the sum-to-zero contrasts in order to get the Type III ANOVA table.

We may notice in the example considered here that the sum of squares as well as the F statistics are identical for the Type I and Type III ANOVA tables. This is because the `TVbo` data is balanced. When the data is unbalanced the results depend on the type of the hypothesis that is used (see Section 2.4 for the discussions).

5.1.5 rand function

`anova` function applied to two nested models with the same fixed structure produces likelihood ratio test on random effects in question. The **lmerTest**

package provides a function with the name `rand` that produces an ANOVA-like output, which contains tests for each random effect of a LMM in question. The function is valuable in situations when there is a large number of random effects, so specifying nested LMM and applying multiple times `anova` function on them becomes too cumberscome. Let us for illustration consider the `TVbo` data and the attribute `Coloursaturation` as a response variable. A model that contain all possible random effects is then the following one:

So 9 random effects are part of the `tv` model. In order to test for significance each of the random effects, a reduced model needs to be constructed and then `anova` function should be applied. Let us, for instance, test the `Assessor` effect. First we construct a reduce model, that does not contain the `Assessor` effect. Then apply the `anova` method to two models. In a similar way another eight reduced models need to be constructed and tested with model `tv`. If, instead, we run the `rand` function on model `tv`, we get the results of the LRT applied to each random effect presented in a compact way:

Table 5.1: Likelihood ratio tests for the random-effects for the TVbo data for attribute Coloursaturation

	χ^2	Chi.DF	p-value
Assessor:TVset	26.22	1	<0.001
Assessor:Picture	0.00	1	1.0000
Assessor:Picture:TVset	2.71	1	0.0995
Repeat	0.00	1	1.0000
Repeat:Picture	0.00	1	1.0000
Repeat:TVset	0.00	1	1.0000
Repeat:TVset:Picture	0.09	1	0.7583
Assessor	0.02	1	0.8846
Assessor:Repeat	2.07	1	0.1506

The `rand` function is also a basis for the simplification of the random structure in the `step` method.

5.1.6 Details on the implementation of the Satterthwaite's approximation in the lmerTest

In the package **lmerTest** the Satterthwaite's approximation to degrees of freedom is implemented for a one-degree-of-freedeom test t test and multi-degree-of-freedom F test. The algorithm that is used is the one proposed by Fai and Cornelius (1996) for the F test and Giesbrecht and Burns (1985) for the t test,

which are described in Section 2.3.1. Most of the estimates, needed for calculating the degrees of freedom can be easily extracted from an `lmer` object via the functions, that the **lme4** package provides (for instance, $\hat{C}$ in Equation 2.8 can be extracted via function `vcov`). However, not everything can be directly obtained via **lme4**: for instance, the asymptotic variance-covariance matrix A Equation 2.10 is not directly extractable from an `lmer` object The asymptotic theory of maximum likelihood (see Serfling (1980)) shows that an asymptotic variance-covariance matrix A of the estimated variance-covariance parameters can be obtained as two times the inverse of the second derivative of the objective function evaluated at the optima. In the **lmerTest** package a function that calculates the deviance (minus 2 times the log-likelihood) is specified for the variance-covariance parameters. This function calls directly the deviance function, which is implemented in the C++ code in the **lme4** package. Since calculation of the hessian involves calling the deviance function multiple times, it is of a great computational benefit to call directly the C++ code in evaluation of the deviance function. The function, that calls directly the C++ code has appeared in the **lmerTest** in 2014, before the function calculating the deviance based on the variance parameters was part of the R code in the **lmerTest** package, which indeed resulted in taking much more time to campute the hessian.

5.2 R package SensMixed

The **SensMixed** was specificallly developed for the sensory practitioners in order to apply/use the methods introduced in chapters 3 and 4. The functions of the **SensMixed** package can be directly used from the R command line or one may use the GUI, that forms part of the package and is implemented via the **shiny** R-package (Chang et al., 2015). The GUI can be called by typing `SensMixedUI()` in the R-Console. The tutorial for the functionalities of the **SensMixed** package and its GUI is in Appendix I. As discussed in Section 3.6.7 it is not straightforward to perform post-hoc analysis for the product difference within a mixed assessor model (MAM) framework. This Section consists of implementation details for performing pairwise product comparisons within MAM.

5.2.1 Implementation of post-hoc in SensMixed

As described in Section 2.5 the test for pairwise comparisons of two products can be performed using the t test. The t test with the Satterthwaite's approximation degrees of freedom as implemented in **SensMixed**:

```
calculateTtest <- function(model, theopt, l, alpha = 0.05){
  ## specifying a function that calculates lCl'
  ##based on theta parameters
  vss <- vcovJSStheta2(model)

  ## evaluate varcor at the theta parameters
  varcor <- vss(t(l), theopt)

  ## calculate standard error
  std.err <- sqrt(varcor)

  ## lsmeans estimate
  estim <- l %*% fixef(model)

  #calculate the t statistics
  t.stat <- estim / std.err

  ## Satterthwaite's approximation to degrees of freedom
  df1 <- calSatterthDenom(model, theopt, l, vss)

  #calculate p value
  p.value <- 2 * (1 - pt(abs(t.stat), df = df1))


  #calculate CIs
  ci.low  <- estim - abs(qt(alpha/2, df1)) * std.err
  ci.upp <- estim + abs(qt(alpha/2, df1)) * std.err

  ## return a list containing t statistics and standard error
  return(list(estim = estim, std.err = std.err, p.value = p.value,
              ci.low = ci.low, ci.upp = ci.upp))
}
```

`calculateTtest` takes the following arguments: `model` an LMM model, `theopt` a vector of estimates of variance-covariance parameters, `l` a contrast vector and `alpha` a Type 1 error rate (default one is 0.05). First a variance-covariance matrix of fixed effects as a function of `theopt` is specified via `vcovJSStheta2` . Then the variance-covariance of an estimate of parameter in question (specified via vector l) is evaluated at the optima and the standard error is calculated. Then the t statistics is calculated as specified in Equation 3.12. Then the function `calSatterthDenom` is called, that calculates the degrees of freedom based via Satterthwaite's method of approximation Giesbrecht and Burns (1985). Finally, the confidence intervals are calculated as specified in Equation 2.14. The

calculation of Satterthwaite's degrees of freedom is implemented as follows:

```
calSatterthDenom <- function(model, theopt, l, vss){
  dd <- devfun5(model,  getME(model, "is_REML"))
  h <- hessian(dd, theopt)
  A <- 2*solve(h)

  g <- grad(function(x)  vss(t(l), x), theopt)
  denom <- t(g) %*% A %*% g
  varcor <- vss(t(l), theopt)
  df1 <- 2*(varcor)^2/denom

  ## return denominator degree of freedom
  return(df1)
}
```

Here the `devfun5` is the deviance function. Note, that here the asymptotic variance covariance matrix A is calculated as a 2 times the inverse of the hessian of the deviance function (for the discussions about that see Section 5.1.6).The rest of the code of the `calSatterthDenom` function simply replicates the algorithm described in Section 2.3.1

Here I compare product pairwise differences for the model with scaling correction to the one without. For illustration, I consider the `TVbo` data and the `Coloursaturation` attribute as a response variable. A 2-way model with one `product` effect is specified as follows:

```
tv <- lmer(Contrast~ product + (1|Assessor) +
             (1|Assessor:product) , data=TVbo)
```

Then, the arguments needed for the `calculateTtest` are obtained from model `tv`. The contrast vector is specified for comparing the first with the fourth product, which corresponds for comparing Picture 4 to Picture 1. Finally, the `calculateTtest` is called in the following code:

```
sigma.m <- sigma(tv)
thopt <- getME(tv, "theta")
theopt <- c(thopt, sigma.m)
l1 <- c(0, 1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0)
calculateTtest(tv, theopt, l1)


## $estim
```

```
##       [,1]
## [1,] -1.86
##
## $std.err
##      [,1]
## [1,] 0.65
##
## $p.value
##         [,1]
## [1,] 0.00537
##
## $ci.low
##       [,1]
## [1,] -3.16
##
## $ci.upp
##        [,1]
## [1,] -0.568
```

Similarily, I specify MAM, get the estimates of relative variance-covariance parameters as well as of σ and call `calculateTtest` function:

```
TVbo$x <- scale(ave(TVbo$Coloursaturation, TVbo$product), scale = FALSE)
lmTV <- lmer(Contrast~ product +   Assessor:x  + (1|Assessor) +
               (1|Assessor:product) , data=TVbo)
sigma.m <- sigma(lmTV)
thopt <- getME(lmTV, "theta")
theopt <- c(thopt, sigma.m)
calculateTtest(tv, theopt, l1)
```

```
## $estim
##       [,1]
## [1,] -1.86
##
## $std.err
##       [,1]
## [1,] 0.556
##
## $p.value
##         [,1]
## [1,] 0.00125
```

```
##
## $ci.low
##       [,1]
## [1,] -2.97
##
## $ci.upp
##        [,1]
## [1,] -0.755
```

From the output it is clear that the p value for the test of product differences is less when considering MAM (the variance-covariance parameters from MAM) than for the model without the scaling correction. Also the confidence interval is narrower for the MAM. This shows that indeed the test for the product differences become more powerful when accounting for the scaling effect through the MAM framework.

Chapter 6

Concluding remarks

Automated model building approach One could be concerned about the "inflation of p-values" as a consequence of some of the selection procedures put forward in paper in Appendix A. On the other hand, the point in favour to the methodology presented there usually would be that in complicated settings, in practice the full model analysis might never be done, and one would rely on simple models discussed in this chapter. It is pretty obvious that if e.g. a fixed higher order interaction effect is never tested, it will never be detected by a too simplistic analysis - so the power to detect it in such a case is basically zero. Maybe one could show how a certain amount of such cases would render (some of) the "selection-based" approaches definitely better than the simplistic methods. But what is the actual consequence of never identifying an important higher-order random effect? And how severe is the inflation of p-values really in various settings? We leave all that for the future research.

The lmerTest package The **lmerTest** package has been widely used for the last few years (around 12000 installations yearly). Hence, we recognize the need to maintain stability of the package so that it continues to be broadly useful. At the moment the `lmer` function of the **lme4** package does not explicitly support models with residual error structures, like AR(1) or other time/spatially dependent error structures often employed for longitudinal modelling and data analysis. But work is going on in this direction. **flexLambda** branch of **lme4**

on Github https://github.com/lme4/lme4/tree/flexLambda is intended to allow a much wider variety of models (e.g. auto-regression). It will hence be a really nice further development to extend the **lmerTest** package similarly. Particularly interesting is then that the Satterthwaite approach for approximating the degrees of freedom for fixed effect F-testing implemented by us in **lmerTest** as a fast and good alternative to the likelihood ratio test and would still be a feasible approach for such models. And investigations of such performance issues could be part this research. The combination of handling big data and mixed modelling is given special focus in research as the computational challenges for big data render standard implementations not feasible for big data. The lme4 package uses sparse matrix techniques and was developed exactly for being able to handle big data in terms of a high number of observations. Computation time is still an important issue in **lmerTest**. Even though the Satterthwaite's method of approximation is generally faster than Kenward-Roger's, still it is significantly slower than the LRT. Hence further improvements in the code and techniques are needed - I leave them for the future work.

Mixed assessor models and extensions With the help of the **lme4** package together with the **lmerTest** mixed assessor models and their extended versions can be easily constructed and analyzed. The examples from the Section 3.6.7 have shown that considering the scaling effect for the extended versions of MAM make the tests for the product effects more powerful, as is supposed to be (Brockhoff et al., 2015). Performing post-hoc analysis is an essential part of the analysis of sensory studies. It could be a valuable information to know which products do differ between each other. However, for MAM it is not trivial to get the pairwise comparisons between the products. I present the "work-around" approach to perform post-hoc together with the tool facilitating the approach. The approach can also handle extended versions of MAM as well as unbalanced data. By this more powerful tests for the product difference can be obtained.

The MAM that is presented in Brockhoff et al. (2015) and extended versions that are presented in Section 3.10 and in the paper in Appendix E are actually the approximate linear versions to the more correct multiplicative model, that is also presented in (Brockhoff et al., 2015). An important benefit of these approximate approaches is the immense simplification of theory and computations, with only a modest change in interpretation. However, even if the approximate approach is a good one due to many reasons considered here, still it is important to being able to handle also a more complicated multiplicative model, that is being able to handle such models and to perform inference on parameters.

Visualizing results Throughout this project an attention has been given to provide tools that are easy-to-use. Both ConsumerCheck and **SensMixed** apart from providing novel tools to analyze consumer and sensory data, provide graphical user interface, where all the results can be presented visually in a very nice way. The **SensMixed** can be considered as an extended version of ANOVA mixed modelling provided by PanelCheck, since it can account for more complex error structures, multi-way product structures, handling unbalanced data and modelling scaling effects. An important issue in **SensMixed** is the computational time. It can take minutes to run analysis in **SensMixed** compared to PanelCheck where it takes only few seconds. The reason is connected first, that in the automated analysis that the **SensMixed** performs a number of times the mixed effects are constructed and compared between each other, which for large models and big data can be quite computationally intensive. Another issue is the Satterthwaite's approximation to degrees of freedom, that is used in tests for the fixed effects - since it involves calculation of the hessian of the likelihood at the optima (see Section 2.3.1), it becomes also quite computationally intensive in situations with big data and/or complex models.

Bibliography

Bates, D., M. Maechler, B. Bolker, and S. Walker (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-4.

Bavay, C., P. B. Brockhoff, A. Kuznetsova, I. Maitre, E. Mehinagic, and R. Symoneaux (2014). Consideration of sample heterogeneity and in-depth analysis of individual differences in sensory analysis. *FOOD QUALITY AND PREFERENCE 32*, 126–131.

Bavay, C., R. Symoneaux, I. Maitre, A. Kuznetsoya, P. B. Brockhoff, and E. Mehinagic (2013). Importance of fruit variability in the assessment of apple quality by sensory evaluation. *POSTHARVEST BIOLOGY AND TECHNOLOGY 77*, 67–74.

Beck, T. K., S. Jensen, G. K. Bjoern, and U. Kidmose (2014). The masking effect of sucrose on perception of bitter compounds in brassica vegetables. *JOURNAL OF SENSORY STUDIES 29*(3), 190–200.

Brockhoff, P. (1998). Assessor modelling. *FOOD QUALITY AND PREFERENCE 9*(3), 87–89.

Brockhoff, P. B., P. Schlich, and I. Skovgaard (2015). Taking individual scaling differences into account by analyzing profile data with the mixed assessor model. *Food Quality and Preference 39*, 156–166.

Brockhoff, P. M. and I. M. Skovgaard (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference 5*(3), 215–224.

by Littell, O., Milliken, Stroup, Wolfinger, modifications by Douglas Bates, M. Maechler, B. Bolker, and S. Walker (2014). *SASmixed: Data sets from "SAS System for Mixed Models"*. R package version 1.0-4.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2015). *shiny: Web Application Framework for R*. R package version 0.11.1.

Dunn, O. (1961). Multiple comparisons among means. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 56*(293), 52.

Fai, A. H. and P. L. Cornelius (1996). Approximate f-tests of multiple degree of freedom hypotheses in generalised least squares analyses of unbalanced split-plot experiments. *Journal of statistical computation and simulation 54*, 363.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance.

Foundation, P. S. Python language reference, version 2.7.

Fox, J. and S. Weisberg (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage.

Giesbrecht, F. and J. Burns (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *BIOMETRICS 41*(2), 477–486.

Gustafsson, A., A. Herrmann, and E. Huber, Frank (2003). *Conjoint measurement : Methods and applications*. Springer.

Halekoh, U. and S. Højsgaard (2014a). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software 59*(9), 1–30.

Halekoh, U. and S. Højsgaard (2014b). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software 59*(9), 1–30.

Harvey, W. R. (1975). Least-squares analysis of data with unequal subclass numbers.

Højsgaard, S., U. H. with contributions from Jim Robison-Cox, K. Wright, A. A. Leidi, and others. (2014a). *doBy: Groupwise statistics, LSmeans, linear contrasts, utilities*. R package version 4.5-13.

Højsgaard, S., U. H. with contributions from Jim Robison-Cox, K. Wright, A. A. Leidi, and others. (2014b). *doBy: Groupwise statistics, LSmeans, linear contrasts, utilities*. R package version 4.5-13.

Jaeger, S. R., L. H. Mielby, H. Heymann, Y. Jia, and M. B. Frøst (2013). Analysing conjoint data with ols and pls regression: a case study with wine. *Journal of the Science of Food and Agriculture 93*(15), 3682–3690.

Kenward, M. and J. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *BIOMETRICS 53*(3), 983–997.

Kenward, M. G. and J. H. Roger (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis 53*, 2583–2595.

Kuznetsova, A., P. Bruun Brockhoff, and R. Haubo Bojesen Christensen (2013a). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. R package version 2.0-0.

Kuznetsova, A., P. Bruun Brockhoff, and R. Haubo Bojesen Christensen (2013b). *SensMixed: Mixed effects modelling for sensory and consumer data*. R package version 2.0-6.

Kuznetsova, A., R. H. Christensen, C. Bavay, and P. B. Brockhoff (2015). Automated mixed {ANOVA} modeling of sensory and consumer data. *Food Quality and Preference 40, Part A*(0), 31 – 38.

Langsrud, y. (2003). Anova for unbalanced data: Use type ii instead of type iii sums of squares. *Statistics and Computing, Stat. Comput 13*(2), 163–167.

Lawless, H. T. and H. Heymann (1997). Sensory evaluation of food: principles and practices.

Lawless, H. T. and H. Heymann (2010). Sensory evaluation of food. 2nd edition.

Lenth, R. V. and M. Hervé (2014). *lsmeans: Least-Squares Means*. R package version 2.13.

MacFie, H. (2007). Consumer-led food product development.

Macnaughton, D. B. (2009). Which sums of squares are best in unbalanced analysis of variance?

McEwan, J. A. (1996). Preference mapping for product optimization. *Data Handling in Science and Technology, Data Handl. Sci. Technol 16*(C), 71–102.

Naes, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference 2*(3), 187–199.

Naes, T., P. B. Brockhoff, and O. Tomic (2010). Statistics for sensory and consumer science.

Nofima Mat, Ås, N. (2008). Panelcheck software.

Peltier, C., P. B. Brockhoff, M. Visalli, and P. Schlich (2014). The mam-cap table: A new tool for monitoring panel performances. *Food Quality and Preference, Food Qual. Preference 32*, 24–27.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-120.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-plus*. Springer Verlag New York, LLC.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romano, R., P. B. Brockhoff, M. Hersleth, O. Tomic, and T. Naes (2008). Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. *Food Quality and Preference 19*(2), 197–209.

SAS (1978). Tests of hypotheses in fixed-effects linear models. Technical report, SAS Institute Inc.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *BIOMETRICS BULLETIN 2*(6), 110–114.

Schaalje, G. B., J. B. McBride, and G. W. Fellingham (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models.

Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. *Data Handling in Science and Technology, Data Handl. Sci. Technol 16*(C), 259–306.

Searle, S. R. (1987). *Linear models for unbalanced data*. Wiley.

Self, S. and K. Liang (1987). Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 82*(398), 605–610.

Senn, S. (2007). *Statistical issues in drug development electronic resource*. John Wiley and Sons.

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. John Wiley.

Speed, F. M., R. R. Hocking, and O. P. Hackney (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association 73*(361), 105.

Steel, R., J. Torrie, and D. Dickey (1997). *Principles and procedures of statistics : A biometrical approach*. McGraw-Hill.

Stram, D. and J. LEE (1994). Variance-components testing in the longitudinal mixed effects model. *BIOMETRICS 50*(4), 1171–1177.

Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika 42*(2), 267–276.

Tomic, O., C. Forde, C. Delahunty, and T. Naes (2013). Performance indices in descriptive sensory analysis - a complimentary screening tool for assessor and panel performance. *FOOD QUALITY AND PREFERENCE 28*(1), 122–133.

Tomic, O., G. Luciano, A. Nilsen, G. Hyldig, K. Lorensen, and T. Næs (2009). Analysing sensory panel performance in a proficiency test using the panelcheck software. *European Food Research and Technology, Eur. Food Res. Technol 230*(3), 497–511.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *BIOMETRICS 5*(2), 99–114.

Weisberg, S. (1985). *Applied linear regression.* John Wiley.

Wiley, D. (1962). The analysis of variance - scheffe,h. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT 22*(3), 627–630.

Appendix A

Automated mixed ANOVA modeling of sensory and consumer data

Kuznetsova, Alexandra, Rune Haubo Bojesen Christensen, Cecile Bavay, and Per Bruun Brockhoff. 2015. "Automated Mixed ANOVA Modeling of Sensory and Consumer Data." Food Quality and Preference 40: 31–38. doi:10.1016/j.foodqual.2014.08.004.

Food Quality and Preference 40 (2015) 31–38

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Automated mixed ANOVA modeling of sensory and consumer data

Alexandra Kuznetsova [a,*], Rune H.B. Christensen [a], Cecile Bavay [b], Per Bruun Brockhoff [a]

[a] DTU Compute, Statistical section, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark
[b] Groupe ESA, UPSP GRAPPE 55, rue Rabelais BP30748, 49007 Angers, Cedex 01, France

ARTICLE INFO

Article history:
Received 30 January 2014
Received in revised form 24 June 2014
Accepted 15 August 2014
Available online 27 August 2014

Keywords:
Mixed-effects models
Automated model building
R program
Conjoint
Consumer preference
ANOVA

ABSTRACT

Mixed effects models have become increasingly prominent in sensory and consumer science. Still applying such models may be challenging for a sensory practitioner due the challenges associated with the choosing the random effects, selecting an appropriate model, interpreting the results. In this paper we introduce an approach for automated mixed ANOVA/ANCOVA modeling together with the open source **R** package lmerTest developed by the authors that can perform automated complex mixed-effects modeling. The package can in an automated way investigate and incorporate the necessary random-effects by sequentially removing non-significant random terms in the mixed model, and similarly test and remove fixed effects. Tables and figures provide an overview of the structure and present post hoc analysis. With this approach, complex error structures can be investigated, identified and incorporated whenever necessary. The package provides type-3 ANOVA output with degrees of freedom corrected *F*-tests for fixed-effects, which makes the package unique in open source implementations of mixed models. The approach together with the user-friendliness of the package allow to analyze a broad range of mixed effects models in a fast and efficient way. The benefits of the approach and the package are illustrated on four data sets coming from consumer/sensory studies.

Introduction

Mixed models are used extensively for analyzing sensory and consumer data. Sensory quantitative descriptive analysis (QDA) data are typically analyzed attribute by attribute using analysis of variance (ANOVA) techniques to extract the important attribute-wise product difference information (Lawless & Heymann, 2010). The proper analysis will typically evaluate the statistical significance of product differences by using the assessor-by-product interaction as error structure (Lawless & Heymann, 2010). This is, what generally in statistics is called a mixed model as both fixed-effects (product differences) as random-effects (assessor differences and assessor-by-product interactions) are present in the modeling and analysis approach. Incorporating random consumer effects for the analysis of e.g. consumer preference data or data from conjoint experiments is on one hand necessary to obtain the proper conclusions from such data and on the other hand similarly leads to mixed models. In the simplest of cases a mixed-effects model (mixed model) analysis can be handled by simple averaging combined with the use of the proper error term coming from a simple ANOVA decomposition of the data. Two often occurring examples of this situation are:

1. Complete consumer preference data with just a single product factor, that is, just a collection of different products coded in a single variable (as opposed to a multifactorial setting), calling for a two-way (block) ANOVA, where the error term is simply the residual error.
2. Complete sensory profile data similarly with just a single product factor, calling for either a 2-way or 3-way ANOVA mixed model depending on the presence of a blocking (replication) factor such as session or product batch. And hence calling for using either the panelist-by-product mean square as the error term or a combination of this with blocking-by-product (Næs, Brockhoff, & Tomic, 2010).

These cases are exactly those covered by the open source software package PanelCheck (Mat & Ås, 2008). However, these simple approaches of analysis have their limitations. With missing values or with more complex study designs one would often benefit from a more detailed analysis. The PanelCheck tool can still be a valuable tool in that using missing values imputation and considering all products together it will in most cases be able to provide some relevant ANOVA information for the situation at hand, and by the

* Corresponding author.
E-mail address: alku@dtu.dk (A. Kuznetsova).

http://dx.doi.org/10.1016/j.foodqual.2014.08.004
0950-3293/

way the by-attribute ANOVA is only a small part of what PanelCheck has to offer. The more detailed univariate analysis of variance provided in this paper becomes relevant in the sensory context in the following example list of situations.

- Unbalanced sensory profile data (for example due to missing observations).
- Incomplete consumer preference data.
- 2-(or higher) way product structure in sensory profile data. (Beck, Jensen, Bjoern, & Kidmose, 2014).
- 2-(or higher) way product structure in consumer preference data (Conjoint) (Jaeger, Mielby, Heymann, Jia, & Frost, 2013).
- Extending Conjoint to include consumer background/design variables or factors/covariates.
- Complex blocking, product replication, product batch structures in as well sensory as consumer preference data.
- Extending external preference mapping to include product and consumer background/design factors/covariates.

Even though commercial and open source software exist for the relevant mixed modeling in such situations, it maintains to be a challenge to apply mixed models for a sensory practitioner. The questions arise as to which model to consider, which variables to include and what the interpretations of the results are. We have developed an **R** package named lmerTest (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2013) that will help answer these questions. Moreover, it will do almost all the model selection work for the practitioner and present the results of the model selection process together with post hoc analyses in a nice and user-friendly way.

The paper is organized such that first, in Section 2, we define basic mixed-effects models, then in Section 3 we discuss why mixed-effects models are important for sensory profile and consumer data and introduce aspects of the model building approaches. In Section 4 we present the lmerTest package and the details of automated mixed modeling, and in Section 5 give four examples showing the usefulness of the automated mixed modeling together with the package in the situations mentioned above. The paper ends with discussions and conclusions in Section 6.

Theory: basic mixed models

Let us consider a simple example of a sensory experiment where we have I assessors, J products and R replicates. This type of data can be described by a mixed ANOVA for replicated two-way data, where as effects we have factors A (assessor) and B (product). A reasonable model can then be written as

$$y_{ijr} = \mu + a_i + \beta_j + d_{ij} + \epsilon_{ijr} \quad (1)$$

where a_i and β_j are main effects for factors A and B and d_{ij} is the effect corresponding to interaction between A and B. If we consider the effects of factor A random, then this implies that the effects a_i (assessor) and d_{ij} (interaction between assessor and product) are random:

$$a_i \sim N(0, \sigma^2_{assessor})$$
$$d_{ij} \sim N(0, \sigma^2_{assessor \times product})$$
$$\epsilon_{ijr} \sim N(0, \sigma^2_{error})$$

where all the random-effects are independent of each other.

A commonly used test statistic for fixed effects hypotheses is F-test statistic. For complex mixed models, e.g. with unbalanced data sets, the F-test statistic will generally not be exactly F-distributed. The common approach is to assume that the test statistic approximately follows an F-distribution and calculate an approximate number of denominator degrees of freedom. Two degree of freedom approximations well-known in the statistical literature are Satterthwaite's (Satterthwaite, 1946) and Kenward-Roger's approximations (Kenward & Roger, 1997). Both of these are implemented in the lmerTest package.

Mixed effects model building

When building a mixed effects model a number of questions arise such as which effects to consider as random, which ones as fixed and which effects to include at all.

Considering the assessor effects random is generally regarded by the sensory field as the proper approach (Lawless & Heymann, 2010). The reason to consider them random is based on the interest in the population of assessors rather than to specific assessors. This means that we want to know the variation among assessors rather than estimates of effects of each assessor and to be able to properly account for that. So in model (1) we are interested in estimating $\sigma^2_{assessor}$ and $\sigma^2_{assessor \times product}$. Moreover, Næs and Langsrud (1996) showed that in situations with interactions between assessors and products, considering assessors as fixed-effects may lead to a conclusion that differences between products are larger than they really are. Therefore the assumption of random assessor effects is usually the most appropriate. For consumer tests, following the same arguments, treating the consumer effect as random is also the most natural. The same goes for the replication/session effect if present.

Having decided on which effects to include as random and which as fixed, the question arises as to which approach of model selection to use. Model selection in general, and selection of regression and ANOVA type models in particular, are controversial topics with many highly opinionated papers in the statistical literature (Jiang, Rao, Gu, & Nguyen, 2008; Ibrahim, Zhu, Garcia, & Guo, 2011; Fan & Li, 2012; Scheepers, Tily, Levy, & Barr, 2013; Peng & Lu, 2012). A particular challenge for model selection of mixed-effects models is how to handle the two types of effects; random-effects and fixed-effects. If the random effects are not well chosen, this will affect the estimates and the hypothesis tests of the fixed-effects. Vice versa, variation in the response variable not modeled in terms of fixed-effects can partly end up in the random effects. In this paper we take a rather heuristic, but practical data-driven approach to the problem and consider the backwards selection approach based on step-wise deletion of model terms with high p-values (Diggle, Heagerty, Liang, & Zeger, 2002; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). In this approach the largest possible model is considered at the first place which includes all possible fixed and random effects that are at least in principle supported by the design. Then the simplification of the random structure is performed. Finally the fixed effects are also incrementally eliminated following the principle of marginality, that is the effects that are contained in any other effects are retained in the model when the effects that they are contained in are found to be significant according to the specified Type 1 level. Lower order interactions are contained in the higher order interactions, so when the higher order interactions are found to be significant, the lower order interactions are kept in the model. The marginality principle is used to enhance interpretability of the various fixed effects and tests thereof. The random effects are part of the overall covariance structure and there is no tradition nor reason for applying a similar principle for these effects. The most important random effects should be included to model the variance structures as good as possible. Even it could be quite meaningful to allow for the pooling effect that would be the consequence of eliminating a random main effect while keeping a random

interaction effect with this factor. In the lmerTest package we have implemented this approach, that simplifies the model automatically. Details of this algorithm are given in the next session. The flexibility of simplification of a mixed model by having different options in the lmerTest package makes the approach useful in a broad range of situations. Indeed the researchers may argue that in models with random slopes the random part should not be simplified (Scheepers et al., 2013). With the lmerTest a practitioner may choose to simplify or not the random part and which Type 1 level to use. The same goes for the fixed part.

The automated model selection presents an important development not only in a general statistical context, but for the analysis of sensory and consumer data in particular. Often, in our field, it maintains to be a challenge to apply mixed models, and a substantial statistical expertise is often needed to identify which models should or could be used, how they should be applied and interpreted. An easy finding of a suitable model by the principle of parsimony together with relevant post hoc analyses can be an important tool for the practitioner and lead to the investigation and identification of important effects otherwise completely ignored due to too simplistic initial model choices.

The lmerTest package

Presentation of lmerTest

The lmerTest package (Kuznetsova et al., 2013) builds on top of the lme4 package (Bates, Maechler, Bolker, & Walker, 2013). The lme4 package is probably the most well-known and most used **R** package for fitting mixed-effects models. Some of its main strengths are the userfriendly interface, the ease with which complex random-effects structures are specified and its ability to quickly fit models to large data sets with millions of observations.

It's most serious weakness (with widespread consequences) is that it does not provide *p*-values for parameter estimates and model terms based on *F*-statistics in ANOVA tables.

The lmerTest package extends the lme4 package in a number of ways by providing:

1. Fixed-effects ANOVA table with type-1 and type-3 *F*-tests using Satterthwaite or Kenward-Roger denominator degrees of freedom approximations.
2. *t*-tests for fixed-effects parameter estimates using Satterthwaite degrees of freedom approximations.
3. Automated model selection with backward elimination of random as well as fixed-effects terms.
4. Post-hoc methods to compute population means and pairwise comparisons of factor levels.
5. Plotting features for the post hoc methods.

Tests for fixed-effects and random-effects

Users of other software packages like SAS have been able to obtain *F*- and *t*-tests for fixed-effects terms and fixed-effects parameters for a long time using Satterthwaite and Kenward-Roger approximations to the denominator degrees of freedom. With the lmerTest package such tests are now also available as open source for the **R** users. The Kenward-Roger degrees of freedom approximation uses the implementation in the pbkrtest package (Hjsgaard, 2013), while the Satterthwaite approximation is implemented in our lmerTest package. The calculation of Satterthwaite approximation is less computational intensive than Kenward-Roger. In cases with a moderately large sample size Satterthwaite approximation is considered to be conservative (West, Welch, & Galecki, 2007). For small sample sizes or for models with complex variance structures there are indications that the Kenward-Roger approximation may give more correct results (Schaalje, McBride, & Fellingham, 2002).

Tests of random-effects terms are performed using likelihood ratio tests *LRT* as accurate *F*-tests are unavailable for general mixed-effects models. We suggest to use a doubled α level for the significance level α for the random part (e.g. in case one would like to test with the significance level 0.05, then one should take $\alpha = 0.1$). This is due to the fact that changing from the more general model to the more specific model involves setting the variance of certain components of the random-effects to zero, which is on the boundary of the parameter region, hence asymptotic results for *LRT* (*LRT* asymptotically follows χ^2_1-distribution with one degree of freedom) have to be adjusted for boundary conditions. Following self and Liang (1987) and Stram and Lee (1994) the *LRT* more closely follows an equal mixture of χ^2-distributions with zero degrees of freedom (a point mass distribution) and one degree of freedom. The *p*-value from this test can be obtained by halving the *p*-value from the test assuming *LRT* follows χ^2_1. In the package, the α level for the analysis of the random part by default is set to 0.1, while for the fixed part the α level is set to 0.05.

Automated model selection

Selection of an appropriate mixed-effects model can be described as a process involving three steps: (1) specification of the mixed-effects model, (2) simplification of the random-effects structure, and (3) simplification of the fixed-effects structure. The lmerTest package facilitates the last two steps, while the user is still responsible for the first step. In the following these three steps are described in further detail.

Step 1: *Specification of the full model*. Specify a model M where the fixed and random parts contain all explanatory variables and as many interactions as possible. It may not be possible to specify all the variables and interactions that one may think of due to restrictions in the experimental design. An example could be when number of levels for some high order interaction term could be equal to the number of observations in the data, so there is not enough data to estimate the term. We suggest that a selection of variables a priori thought most likely to contribute be specified as part of the initial model.

Step 2: *Analysis of random-effects*.

1. For each random-effect r_i in M do:
   (a) Create a reduced model M_i by eliminating r_i from M
   (b) Calculate p_i, the p-value from the likelihood ratio test of comparing M to M_i
   (c) Save p_i and M_i
2. Find p_{max}; the maximum of all p_i and let M_{max} denote the corresponding model, that is the model without the effect corresponding to p_{max}.
3. Set M to M_{max}. If p_{max} is higher than α level then go back to 1, otherwise stop.

Model M is the output of the algorithm, and we save it for *Step* 3. If in *Step* 1 the random part of M contains slopes (random-coefficient model), then the principle of simplification of such random effects is similar – the effect that contains slopes and intercept is incrementally reduced by removing first non-significant slopes and then non-significant intercepts. The details may be found in the manual (Kuznetsova et al., 2013). In Section 5.3 there is an example of the automated analysis of such a random-coefficient model, the code for the analysis is given in Section Appendix.

Step 3 : *Analysis of fixed-effects*

1. Construct an ANOVA table for M, calculate F-statistics and p-values for each fixed-effects term.
2. Consider the effects that are not contained in any other effects in M. The effect with the highest p-value (p_{eff}) is identified and a model without this effect M_{eff} is constructed.
3. Set M_{eff} to M. If p_{eff} is less than α level or if there are no more fixed-effects then stop, otherwise go to 1.

The default α level in *Step* 2 is twice bigger than in *Step* 3. Model M from *Step 3* is the final model selected by the algorithm.

Some examples and explanations of how to perform model selection with the lmerTest package are given in the following sections.

Improved sensory analysis

In this section we present four examples of how automated modeling can improve sensory analyses.

- Multiway product structures in sensory profile data.
- Conjoint analysis with consumer background variables.
- External preference mapping with product and consumer background variables.
- Unbalanced sensory profile data with product, replication and batch structures.

The analysis was performed using the lmerTest package. The **R**-code of the first three examples is given in the Appendix. Only data for the last example are not available in the lmerTest package.

Multiway product structures in sensory profile data

In this example we consider a dataset on tests of TV sets produced by the highend HIFI company Bang and Olufsen A/S, Struer, Denmark, and was used for a workshop at the 8th Sensometrics Meeting in Norway in 2008. The data are available in lmerTest as `TVbo`.

The main purpose in this study was to assess 12 "products", specified by two features: Picture (factor with 4 levels) and TVset (factor with 3 levels). So all in all 12 "products" in 2 replications were assessed by 8 assessors for 15 different response variables. In this example we choose the attribute colourbalance as our response variable of interest.

To specify an initial model (*Step 1*), we consider assessor and replication effects as random (for discussions see Section 2), TVset and Picture effects as fixed. The initial model that incorporates all possible interactions would then be the following one:

$$Y = t + p + tp + A + R + TA + PA + TR + PR + TPA + TPR + \epsilon \qquad (2)$$

Here t, p, A and R correspond to TVset, Picture, assessor and replication factors accordingly, Y correspond to attribute colourbalance. Small letters represent fixed effects, capital letters represent random effects. Combination of the letters represent interaction, so e.g. TA means random effect for interaction between TVset and assessors. Model (2) is the largest model.

Table 1 and Table 2 present *Step 2* and *Step 3* of the automated analysis of the random and fixed-effects respectively of our initial model (2). The effects that have kept in the elim.num column are the ones that form the simplest plausible model according to the principle of parsimony given by the default type I levels ($\alpha = 0.10$ for the random-effects and $\alpha = 0.05$ for the fixed-effects).

According to Table 1 there are three significant random-effects that enter our simplified model. Table 1 provides the information about the error structure used for the fixed-effects together with interpretable information of its own rights: "products" interact with assessors but most of the interaction is related to the TVset effect rather than the Picture effect (p-value for the interaction between assessors and TVset is less than 0.001 and p-value for the interaction between assessors and Picture is 0.077). And similarly for the interaction effects between "products" and replication: it is purely a TVset related effect.

From Table 2 it can be seen that there is an interaction between TVset and Picture. The two-way interaction plot in Fig. 1 confirms that. From Fig. 2 it can be seen that the most different are "products" with the different levels for TVset: TV 2 and TV 3. The plots and tables are the ones provided by the lmerTest package.

The parsimonious model that the function provided by the lmerTest identified has quite a complex random structure and a multiway product structure in the fixed part. Considering right away a simple model with a one-way product structure (combination of TVset and Picture, which would result in one fixed product effect) and a simple random structure (e.g. considering just the assessor effect) would not provide this valuable insight into the data.

Conjoint analysis with consumer background variables

The data is a rating-based conjoint-type study and comes from Næs, Lengard, Johansen, and Hersleth (2010). Four different hams were analyzed and compared with each other. The information about origin of the products (factor with 2 levels) was given in such way that both correct and incorrect information was presented to the consumers. Therefore the study was a $4*2$ design. 81 consumers gave the liking scores to the products. Two consumer background information variables were available: sex as a factor and age as a continuous variable. The data are available in lmerTest as ham.

To specify an initial model (*Step 1*), we choose consumer as a random-effect, this implies interaction between products and consumers and interaction between information and consumers to enter the model as random-effects. The three-way interaction between information, products and consumers is not part of the initial model since the number of levels for this effect is equal to the number of observations in the data, so there is not enough data to estimate it. The initial model that incorporates all possible interactions would then be:

$$Y = p + i + s + a + pi + is + sa + pa + ps + ia + pis + isa + psa + pia + pisa + C + PC + IC + \epsilon \qquad (3)$$

Here p, i, s, a and C correspond to products, information, sex, age and consumers accordingly, Y correspond to the liking scores of consumers.

Tables 3 and 4 present *Step 2* and *Step 3* of the automated analysis of the random and fixed-effects respectively of our initial model.

Table 1
Likelihood ratio tests for the random-effects and their order of elimination representing *Step 2* of the automated analysis for the TVbo data.

	χ^2	Elim.num	p-Value
TVset × Picture × replication	0.00	1	1.000
Replication × Picture	0.00	2	1.000
Assessor × TVset × Picture	0.00	3	1.000
Replication	0.10	4	0.748
Assessor	0.32	5	0.572
Assessor × TVset	69.52	Kept	<0.001
Assessor × Picture	3.12	Kept	0.077
Replication × TVset	4.99	Kept	0.025

Table 2
F-tests for the fixed-effects and their order of elimination representing *Step 3* of the automated analysis for the TVbo data.

	F	Elim.num	*p*-Value
TVset	3.90	Kept	0.042
Picture	1.37	Kept	0.281
TVset × Picture	4.47	Kept	<0.001

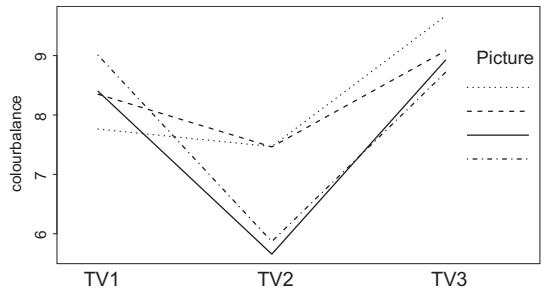


Fig. 1. 2-way interaction plot for TVset and Picture for the TVbo data.

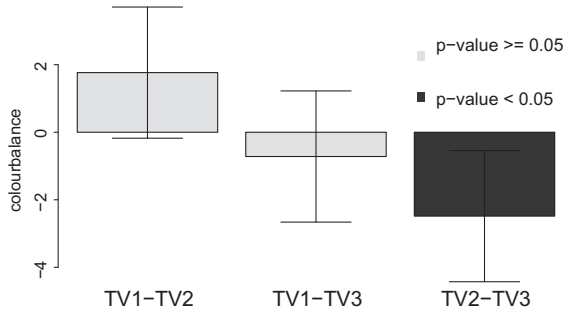


Fig. 2. Barplots for differences of population means for TVset effects together with 95% confidence intervals for the TV data.

It can be seen that model (3) was significantly reduced (see Table 3 and Table 4): only two random-effects and two main fixed effects are part of the final parsimonious model. From Table 3 we see that there is no significant interaction between information effect and consumers. Consumers interacts with products, meaning that consumers differ in their liking of ham products. From Table 4 age appears not to have an impact on consumer's choice of products, similarly sex does not have an impact on consumer's choice. In this example we could have chosen right away a simple model without considering background information of consumers. But then there would be no justification of not including them in the model. The automated mixed modeling that the lmerTest package provides gives some evidence (according to significance level α=0.05) to the fact that these effects are not significant and do not influence liking scores of consumers. Tables are those provided by the lmerTest package.

Table 3
Likelihood ratio tests for the random-effects and their order of elimination representing *Step 2* of the automated analysis for the ham data.

	χ^2	Elim.num	*p*-Value
Information × consumer	1.62	1	0.203
Consumer	3.09	Kept	0.079
Product × consumer	174.16	Kept	<0.001

Table 4
F-tests for the fixed-effects and their order of elimination representing *Step 3* of the automated analysis for the ham data.

	F	Elim.num	*p*-Value
Product × information × sex × age	1.46	1	0.225
Product × sex × age	0.13	2	0.945
Product × information × sex	1.19	3	0.315
Product × sex	0.18	4	0.907
Product × information × age	1.45	5	0.227
Product × age	0.81	6	0.490
Product × information	2.08	7	0.102
Information × sex × age	3.18	8	0.075
Information × age	0.00	9	0.944
Sex × age	0.71	10	0.401
Age	0.02	11	0.903
Information × sex	0.84	12	0.361
Sex	0.88	13	0.351
Product	3.83	Kept	0.010
Information	3.87	Kept	0.050

External preference mapping with product and consumer background variables

The carrots data comes from The Royal Veterinary and Agricultural University, Denmark, and is an example of external preference mapping. 103 consumers scored their preference of 12 danish carrot types on a scale from 1 to 7. In addition to the consumer survey, the carrot products were profiled by a trained sensory panel, with respect to a number of sensory (taste, odour and texture) properties. The first two principal components in a principal component analysis (PCA) on the product-by-attributes panel average data matrix were extracted (sens1 and sens2). Sens1 mainly measures bitterness versus nutty taste, sens2 measures mainly sweetness. The objective of the study was to relate liking scores across carrot products. We have included also some additional consumer background information: size (factor, indicating the number of family members living in a house), age (factor with 4 levels) The data are available in lmerTest as `carrots`.

To specify an initial model (*Step 1*), we choose consumers and products as random-effects. product effect is considered as random, since we wish to consider the entire population of carrot products instead of only the 12 specific products investigated in this experiment. Interaction between products and consumers does not enter the initial model, since there is not enough data to estimate it (number of levels of this factor is equal to the number of observations in the data). Interaction between sens1 and consumers and interaction between sens2 and consumers are part of the random structure of the initial model and represent two random slopes for each sens1 and sens2 at each level of the consumer effect. So the initial model that we choose is quite a complex ANCOVA random-coefficient model and can be written in the following form:

$$\begin{aligned} Y = {} & a + s + sens_1 + sens_2 + as + sens_1 s + sens_1 sens_2 + sens_1 s \\ & + sens_2 a + sens_2 s + sens_1 sens_2 a + sens_1 sens_2 s + sens_1 as \\ & + sens_2 as + sens_1 sens_2 as + P + C + Sens_1 C + Sens_2 C + \epsilon \end{aligned} \quad (4)$$

Here $a, s, sens_1$, $sens_2, P$ and C correspond to age, size, $sens_1$, $sens_2$, products and consumers accordingly. Y corresponds to the liking scores of consumers. $C + Sens_1 C + Sens_2 C$ represents random intercept and two random slopes for $Sens_1$ and $Sens_2$ accordingly at each level of C. The covariance structure comprises variance for the intercept, variances for each of the slopes and all the correlations. One may or not include correlations between different random effects. The purpose here is not the analysis of the correlation structure. Instead we put all the correlations in order not to make any assumptions. So when the effect is eliminated then the

relevant correlations are eliminated as well. Table 5 and Table 6 present *Step 2* and *Step 3* of the automated analysis of the random and fixed-effects respectively of model (4).

According to Table 5 two significant random effects are part of the final model: interaction between sens2 and consumer and a single product effect. So the products are differently scored and the consumers disagree in liking the products with respect to the attributes characterized by the sens2 principle component.

From Table 6 we see that there are only two significant main fixed-effects: size and sens2. Age does not have an impact on consumers' choices and the attributes that are characterized by the sens1 principal component also do not have much influence on consumers' choices.

Unbalanced sensory profile data with product, replication and batch structures

The data comes from a sensory study of apples (Bavay et al., 2013). Three different apples for 9 sensory attributes were tested. There were all in all 19 assessors and they tested each variety of apples 4 times. Fruits were cut into pieces and each piece got a fruit number, factor with 74 levels. First objective was to verify assessors' ability to discriminate. Second goal was to observe fruit-to-fruit variability and to demonstrate the fact that each individual apple had also an impact on the response, and therefore fruit number should also be present in the model. The data is highly unbalanced, which implies that not all softwares (e.g. PanelCheck) are able to construct models for it. In this example we will do the automated mixed effects model selection using the lmerTest package for each attribute and will create plots that will represent results for all attributes at once. The initial model is the same for each attribute and has the following form:

$$Y = p + A + F + R + PA + PF + PR + AR + AF + FR + PAR + \epsilon \qquad (5)$$

Here p, A, F and R correspond to products, assessors, fruit numbers and replications accordingly. Y corresponds to the scores of an attribute. So products enter as a fixed-effect, assessors, fruit numbers and replications enter as random-effects, the rest are the random effects that represent all possible interactions between random effects and between products and random effects plus an error term. As there is only one fixed effect product we may say that there is only a simplification of the random structure in this example. Fig. 3 represents the sequential chi-squared values (i.e. from the stepwise selection process) for the test of random-effects for each sensory attribute. The random effects that had zero variances for all attributes were not included in the plot. The colors of the bars represent the significance of the effect. It can be seen that the effect corresponding to fruit number is present in all final models for all attributes, therefore we cannot ignore this information while constructing the model. Replications are either negligible, or non-significant (except for the attribute Stiffness). Interaction between assessors and replications is only significant for Acidic attribute. Fig. 4 represents the values of the square root of the F-statistics for fixed-effect product for each sensory attribute. The colors of the barplots represent the significance of the product effect. It can be seen that assessors are able to discriminate the products for all attributes except Mealy and Sweet.

Table 5
Likelihood ratio tests for the random-effects and their order of elimination representing *Step 2* of the automated analysis for the carrots data.

	χ^2	Chi.DF	Elim.num	p-Value
$Sens_1 \times$ consumer	2.23	3	1	0.526
$Sens_2 \times$ consumer	6.37	2	Kept	0.041
Product	13.21	1	Kept	<0.001

Table 6
F-tests for the fixed-effects and their order of elimination representing *Step 3* of the automated analysis for the carrots data.

	F	Elim.num	p-Value
Sens1 × Sens2 × size × age	1.14	1	0.331
Sens2 × size × age	0.39	2	0.760
Sens1 × Sens2 × size	0.57	3	0.452
Sens2 × size	2.01	4	0.160
Sens1 × size × age	1.82	5	0.141
Sens1 × size	0.15	6	0.703
Size × age	1.90	7	0.135
Sens1 × Sens2 × age	2.32	8	0.073
Sens1 × Sens2	0.11	9	0.745
Sens1 × age	0.45	10	0.718
Sens2 × age	0.74	11	0.528
Age	0.21	12	0.886
Sens1	0.52	13	0.488
Sens2	17.48	Kept	0.001
Size	5.63	Kept	0.020

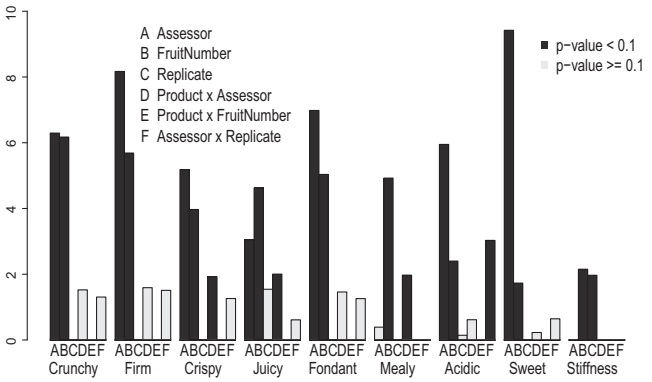


Fig. 3. Barplots for $\sqrt{\chi^2}$ of likelihood ratio test for random-effects for the apples data.

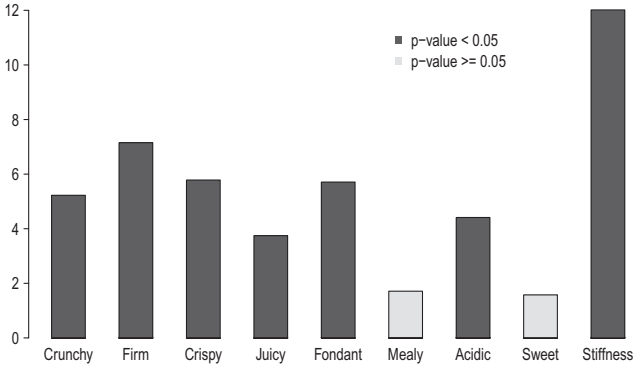


Fig. 4. Barplots for $\sqrt{F}$-statistics for fixed-effect Product for the apples data.

Discussion and conclusion

In this paper we have introduced an approach for automated analysis of mixed-effects models and presented the **R** package lmerTest as an open source implementation of the approach. The aim of the approach is to be able to get important interpretable information from the data and to create a tool that is easy to use for a sensory practitioner. The examples have shown that by using too simplistic models, which for many practitioners may be the choice, some important information might not be found nor accounted for and that the lmerTest package finds. Hence the analysis of sensory/

consumer data can indeed be improved, more insight of the data can be achieved by using the automated analysis implemented in the lmerTest package. The package also provides (open source access to) the *p*-values calculated from *F*-statistics in ANOVA tables with Satterthwaite approximations for denominator degrees of freedom; a novel contribution.

Appendix

R-code for the analysis of the data sets contained in the lmerTest package

First we attach the lmerTest package by typing the following in the R-console.

```
library(lmerTest)
```

Specification of a linear mixed effects model using `lmer` from the lme4 package

All the models are constructed using the same principle. As an example let `response` be the response variable, `eff1`, `eff2` be the main fixed-effects, their interaction would then be `eff1:eff2`, and `eff3` – the random-effect. An `lmer` model to analyze them would then be:

```
modelEx <- lmer(resp ~ eff1*eff2 + (1 | eff3))
```

It can be seen that the model formula consists of two expressions separated by the ~ symbol in the left hand side of the formula the response variable `resp` is specified. The right-hand side consists of one or more terms separated by '+' symbols. The first term is `eff1*eff2`, which represents three fixed effects: two main effects `eff1`, `eff2` and interaction between them `eff1:eff2` (symbol '*' means all main effects plus all possible interactions between them). `(1 | eff3)` is a specification of a random-effect. In general specification of each random-effects term consists of two expressions separated by the vertical bar and enclosed in parentheses. The expression on the right of the vertical bar is a factor (here `eff3`). In a scalar random-effects term (which we have in our case), the expression on the left of the vertical bar is '1'. Such a term generates one random-effect (i.e. a scalar) for each level of the factor. We name the model `modelEx`.

Next we use the `step` function of the lmerTest package in order to analyze/reduce `modelEx` to the parsimonious one.

```
stepEx <- step(modelEx)
stepEx
```

This function provides us with the tables of analysis of the random and fixed parts of the modelEx as well as the tables of post hoc analysis, all that is contained in `stepEx` variable.

Finally we use the function plot in order to plot the results of post hoc analysis.

```
plot(stepEx)
```

Automated analysis of TVbo data set in R

```
  modelTVbo <- lmer(Colourbalance ~ TVset*Picture
+ (1 | Assessor) + (1 | Assessor:TVset)
+ (1 | Assessor:Picture)
+ (1 | Assessor:TVset:Picture)
+ (1 | Repeat) + (1 | Repeat:TVset)
+ (1 | Repeat:Picture), data=TVbo)

stepTVbo <- step(modelTVbo)]

print(stepTVbo)

plot(stepTVbo)
```

Automated analysis of ham data in R

```
  modelHam <- lmer(Informed.liking ~
Product*Information*Gender*Age
+ (1 | Consumer)
+ (1 | Product:Consumer)
+ (1 | Information:Consumer), data=ham)

stepHam <- step(modelHam)

print(stepHam)

plot(stepHam)
```

Automated analysis of carrots data in R

```
  modelCarrots <- lmer(Preference ~
sens2*sens1*Homesize*Age
+ (1 | product) + (1 + sens1 + sens2 | Consumer),
data=carrots)

stepCarrots <- step(modelCarrots)

print(stepCarrots)

plot(stepCarrots)
```

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodqual.2014.08.004

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. r package version 1.0-4. <http://CRAN.R-project.org/package=lme4>.

Bavay, C., Symoneaux, R., Maitre, I., Kuznetsova, A., Brockhoff, P. B., & Mehinagic, E. (2013). Importance of fruit variability in the assessment of apple quality by sensory evaluation. *Journal of Agricultural, Biological, and Environmental Statistics, 77*, 67–74.

Beck, T. K., Jensen, S., Bjoern, G. K., & Kidmose, U. (2014). The masking effect of sucrose on perception of bitter compounds in brassica vegetables. *Journal of Sensory Studies*, 0887–8250.

Diggle, Peter J., Heagerty, Patrick N., Liang, Kung-Yee, & Zeger, Scott L. (2002). *Analysis of longitudinal data*. Oxford University Press.

Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics, 40*, 2043–2068. http://dx.doi.org/10.1214/12-AOS1028.

Hjsgaard, U. H. S. (2013). pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison. r package version 0.3-5. <http://CRAN.R-project.org/package=pbkrtest>.

Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics – Washington, 67*, 495–503.

Jaeger, S. R., Mielby, L. H., Heymann, H., Jia, Y., & Frost, M. B. (2013). Analysing conjoint data with OLS and PLS regression: a case study with wine. *Journal of the Science of Food and Agriculture, 93*, 3682–3690. ISSN:0887–8250.

Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T. (2008). Fence methods for mixed model selection.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects estimators from restricted maximum likelihood. *Biometrics, 53*, 983–997.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). r package version 2.0-0. <http://CRAN.R-project.org/package=lmerTest>.

Lawless, H. T., & Heymann, H. (2010). In *Sensory evaluation of food*. Springer Science+Business Media LLC.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for sensory and consumer science*. John Wiley and Sons Ltd.

Næs, T., & Langsrud, O. (1996). Fixed or random assessors in sensory profiling? *Food Quality and Preference, 9*, 145–152.

Næs, T., Lengard, V., Johansen, S. B., & Hersleth, M. (2010). Alternative methods for combining design variables and consumer preference with information about attitudes and demographics in conjoint analysis. *Food Quality and Preference, 21*, 368–378.

Nofima Mat, N., & Ås (2008). Panelcheck software. <www.panelcheck.com>.

Peng, H., & Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis, 109*, 109–129. http://dx.doi.org/10.1016/j.jmva.2012.02.005.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512–524.

Scheepers, C., Tily, H. J., Levy, R., & Barr, D. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605–610 [01621459, 1537274x].

Stram, D. O., & Lee, J. W. (1994). Variance component testing in the longitudinal mixed effects model. *Biometrics, 50*, 1171–1177.

West, B., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models a practical guide using statistical software*. CRC: Chapman & Hall.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science+Business Media LLC.

Appendix B

lmerTest package: Tests in Linear Mixed Effects Models

Alexandra Kuznetsova, Per B. Brockhoff, Rune H. B. Christensen (2014). lmerTest package: Tests in Linear Mixed Effects Models. (working paper)

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

lmerTest package: Tests in Linear Mixed Effects Models

Alexandra Kuznetsova
Technical University Of Denmark

Per B Brockhoff
Technical University Of Denmark

Rune H.B. Christensen
Technical University Of Denmark

Abstract

One of the frequent questions by users of the mixed model function `lmer` of the **lme4** package has been: how can I get p-values for the F and t tests for `lmer` objects? The **lmerTest** package extends the `lmerMod` class of the **lme4** package, by overloading the `anova` and `summary` functions by providing p-values for tests for fixed effects. We have implemented Satterthwaite's method for approximating denominator degrees of freedom and also the construction of Type I - III ANOVA tables. Furthermore, one may also obtain the `summary` as well as the `anova` table using the Kenward-Roger approximation for denominator degrees of freedom (based on the `KRmodcomp` function from the **pbkrtest** package). Some other convenient mixed model analysis tools such as a `step` method, that performs backward elimination of non-significant effects - both random and fixed, calculation of population means and multiple comparison tests together with plot facilities are provided by the package as well.

Keywords: denominator degree of freedom, Satterthwaite's approximation, ANOVA, R, linear mixed effects models, lme4.

1. Introduction

Linear mixed effects models are tools for modelling continuous correlated hierarchical/multilevel data. During the last decades these models have become more and more prominent in a variety of fields such as physical, biological and social sciences. Various software comercial as well as open-source are capable of fitting these types of models. The focus of this paper is on the open-source R-package **lme4** (Bates, Maechler, Bolker, and Walker 2013). This package

is a well-known and widely used R-package designed to fit linear as well as non-linear mixed effects models. Some of the **lme4** package main strengths are the userfriendly interface, the ability to handle unbalanced data, multiple crossed effects and being very fast even for large data sets.

The `anova` and `summary` functions are two of the main functions providing inference on the parameters of a model. In tests for the fixed effects of a linear mixed effect model, the F-statistics `anova` and the t-statistics `summary` functions are given, though p-values for the corresponding F and t tests are not provided by the **lme4** package. The reason is connected with the fact that generally the exact null distributions for the parameter estimates and test statistics are unknown. So the only way to judge about the significance of the effects is by some sort of approximation and/or simulation based approach. A common way is to use the likelihood ratio test (LRT). This test is fast and is available in the **lme4** package. The downside is that it can produce anti-conservative p-values in a variety of situations, which we discuss in Section 3. A simulation based alternative is the `bootMer` function from the **pbkrtest** package Halekoh and Højsgaard (2014), which is computationally intensive. The authors of the **pbkrtest** package have implemeted the Kenward-Roger's approximation method, which provides accurate p-values, but for some types of models and large data the method could be computationally intensive. Our aim was to provide a method, that is a nice alternative to the widely used LRT. We have implemented Satterthwaite's method Giesbrecht and Burns (1985); Fai and Cornelius (1996) as implemented in SAS software (SAS 1978) and wrapped it into `anova` and `summary` functions for an `lmer` object. We have also integrated the Kenward-Roger's approximation method through the `KRmodcomp` function of the **pbkrtest** package. Hence there are two available alternatives for the `anova` and `summary` methods.

Another contribution of the package is a generation of Type I - III hypothesis contrast matrices that result in producing the corresponding types of ANOVA tables. The Types II and III may be also obtained through the `Anova` function of the **car** package (Fox and Weisberg 2011). However some limitations can be found. For instance, sum-to-zero restrictions on parameters should be used in order to get the correct Type III ANOVA table. In our implementation the generation of three types of the ANOVA tables is invariant with respect to the restrictions used on the parameters of the linear mixed model.

Some other convenience functions such as the `step` function that performs automated elimination of non-significant effects, the `lsmeans` and `difflsmeans` functions that generate the least squares means and the differences of least squares means tables with confidence intervals are provided by the **lmerTest** package. The functions contained in the **lmerTest** package are listed in Table 1.

The paper is structured in the following way: in Sections 2 and 3 we describe the approach taken by Giesbrecht and Burns (1985); Fai and Cornelius (1996) to address the inference problem and compare the approximation methods to the commonly used LR- test. In Section 4 two of the data sets from the **lmerTest** package are introduced. In Section 5 we discuss different types of hypothesis for ANOVA and their implementation in the **lmerTest** package. In Section 6 we introduce least squares means and their implementation in the **lmerTest** package. In Section 7 we introduce our implementation of the step-down model building approach. In Section 8 we describe the methods contained in the package. In Section 9 we discuss the timing issues for approximation methods for a certain class of linear mixed effects models. Section 10 contains discussion and conclusion.

Table 1: Summary of the functions provided by the lmerTest package

Functionalities	`anova`	`summary`	`rand`	`step`	`lsmeans`	`difflsmeans`
output as from **lme4**	✓	✓				
ANOVA-like table for the random effects (LRT)			✓	✓		
Satterthwaite's approximation to degrees of freedom	✓	✓		✓	✓	✓
Kenward-Roger's approximation to degrees of freedom	✓	✓		✓		
Type I, II, III hypothesis tests (SAS-notations)	✓			✓		
Least squares means				✓	✓	
Differences of least squares means				✓		✓
Automated elimination of random and/or fixed effects				✓		

2. Inference and test statistic

A linear mixed model can be specified on matrix form as:

$$y = X\beta + Zu + \varepsilon \quad u \sim N_q(0, G) \quad \varepsilon \sim N_n(0, R) \tag{1}$$

with β representing all fixed-effects parameters, u the random-effects, X the $n \times p$ design matrix for the fixed-effects parameters, and Z the $n \times q$ design matrix for the random-effects. To test hypothesis about the fixed effects β, one may use the LRT. Then a smaller model needs to be constructed with the same error structure as model 1 has:

$$y_0 = X_0\beta_0 + Zu + \varepsilon \tag{2}$$

The LRT statistic for the test of the hypothesis

$$H_0 : \beta \in \Theta_{\beta_0}$$

$$H_1 : \beta \in \Theta_{\beta}$$

where Θ_{β_0} is a subspace of the parameter space Θ_β of the fixed effects β is:

$$T = 2(ll - ll_0)$$

where ll and ll_0 represent the log-likelihoods of models 1 and 2 accordingly. Under the hypothesis, T follows asymptotically a χ^2 distribution. Even though LRT is frequently used, it can produce anti-conservative p-values (Pinheiro and Bates 2000).

One may consider an F test of the hypothesis $H_0 : L\beta = 0$, where L is a contrast matrix of $q = rank(L) > 1$. A test statistic for this hypothesis is:

$$F = \frac{(L\hat{\beta})^\top (L\hat{C}L^\top)^{-1}(L\hat{\beta})}{q} \tag{3}$$

Where $\hat{C}$ is an estimated variance-covariance matrix. Even though the statistic is called F, in general it does not exactly follow an F distribution. A method, known as Satterthwaite, was proposed by Fai and Cornelius (1996) for determining denominator degrees of freedom ν such that: $F \sim F_{q,\nu}$ approximately. We have implemented their work for the F-test and also for a one-degree of freedom test, which corresponds to the t-test with the method, proposed by Giesbrecht and Burns (1985). The details of the algorithm are given in Appendix A. In Kenward-Roger's method the estimated variance covariance matrix $\hat{C}$ is adjusted in order to improve the small sample distributional properties of F and then the Satterthwaite's method-of-moment of approximation is applied. The algorithm may be found in Halekoh and Højsgaard (2014).

3. Comparisons of F tests and LR-tests

As previously mentioned, the LRT can produce anti-conservative p-values. This may occur when the data is unbalanced or when the number of parameters is large compared to the number of observations (Pinheiro and Bates 2000). In Halekoh and Højsgaard (2014) an example where LRT leads to misleading results and where Kenward-Roger's method is accurate is given.

Pinheiro and Bates (2000) provide a simulation study for the LRT based on the `PBIB` data. The `PBIB` data comes from the **SASmixed** package (by Littell, Milliken, Stroup, Wolfinger, modifications by Douglas Bates, Maechler, Bolker, and Walker 2014) and is an example of a partially balanced incomplete block experiment with $i = 1, \ldots, 15$ treatments, $j = 1, \ldots, 15$ blocks and 60 observations. Not every level of treatment appears with every level of blocking factor, but every pair of treatments occur together in a block the same number of times. Pinheiro and Bates (2000) consider the following mixed effects model for this data:

$$y_{ijk} = \alpha_i + b_j + \epsilon_{ijk} \tag{4}$$

$$b_j \sim N(0, \sigma_b^2) \text{ and } \epsilon_{ijk} \sim N(0, \sigma^2)$$

where α stands for a treatment effect, b stands for a random block effect.

In order to compare LRT to the F-test with Satterthwaite and Kenward-Roger approximation methods we performed the same simulation study for a test for a presence of the treatment effect. We performed 1000 simulations from the model with only a random block effect corresponding to the null hypothesis of no treatment effect. The results of the simulations

are represented in Figure 1. It can be seen that the LRT gives for all nominal values anti-conservative p values. It is also clear that both Satterthwaite's and Kenward-Roger's methods p-values are close to the nominal values.

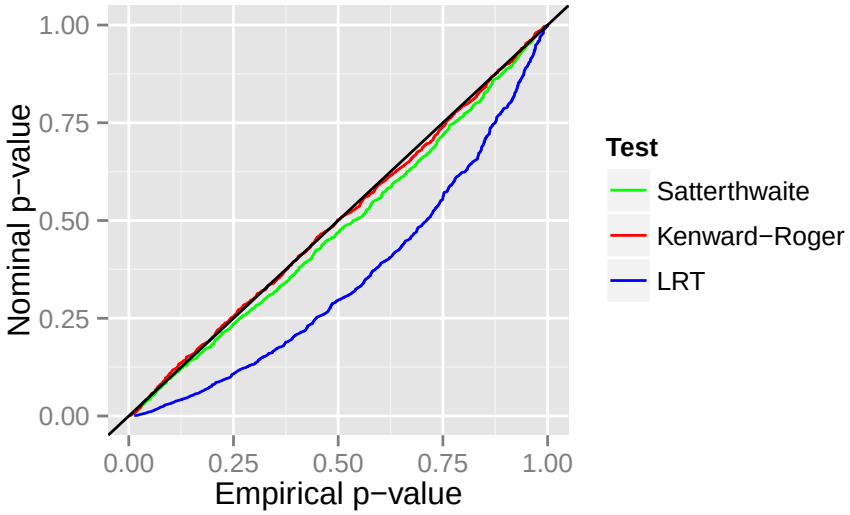


Figure 1: Empirical p values versus nominal p values ranging from 0.001 to 1 for the test of the presence of the treatment fixed effect. The results are based on 1000 simulations from the model with a random block effect applied to the `PBIB` data

4. Data sets

lmerTest includes three data sets from Sensory and Consumer studies. Throughout the paper we will use two of the them: the first one with the name `TVbo` comes from a sensory study and consists of tests of TV sets produced by the highend HIFI company Bang and Olufsen A/S, Struer, Denmark. The second data set is a combination of a sensory and a consumer study and has the name `carrots`.

4.1. The TVbo data

The main purpose in this study was to assess 12 products, specified by two features: `Picture`, a factor with 4 levels and `TVset`, a factor with 3 levels. All in all 12 products in 2 replications were assessed by 8 trained panellists (`Assessor`) for 15 different response variables on a scale from 1 to 14. This type of data is very common in Sensory science (Næs, Brockhoff, and Tomic 2010).

For illustration, let us select the attribute `Sharpnessofmovement` as our response variable. We consider the `Assessor` effect as random since it is generally regarded as the proper approach in the sensory field (Lawless and Heymann 2010). In the fixed part of the model we include `TVset` and `Picture` effects and their interaction. In the random part we also include interaction effects `Assessor:TVset` and `Assessor:Picture`. The choise of including these effects will be later justified in Section 8.4.

A linear mixed effects model for the `Sharpnessofmovement` attribute is then:

$$y_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + c_k + ac_{ik} + bc_{jk} + \epsilon_{ijk} \qquad i = 1, 2, 3 \qquad j = 1, 2, 3, 4 \tag{5}$$

$$c_k \sim N(0, \sigma_c^2) \text{ and } ac_{ik} \sim N(0, \sigma_{ac}^2) \text{ and } bc_{jk} \sim N(0, \sigma_{bc}^2)$$

where α stands for `TVset` effect, β for `Picture` effect, c stands for `Assessor` effect.

4.2. The carrots data

The `carrots` data comes from The Royal Veterinary and Agricultural University, Denmark and is an example of a so-called external preference mapping. 103 consumers scored their preference of 12 danish carrot types on a scale from 1 to 7. In addition to the consumer survey, the carrot products were profiled by a trained panel of tasters, the sensory panel, with respect to a number of sensory properties (taste, odour and texture). The goal was to relate the sensory properties of the products to the consumer liking. Since there was a high number of sensory properties (14), a principal component analysis was performed and two first principle components were extracted that contained most of the information in the sensory properties (`sens1` and `sens2`). `sens1` mainly measured bitterness versus nutty taste, whereas `sens2` measured mainly sweetness. A common method for preference mapping is to fit regression models for the preference as a function of the sensory variables for each individual consumer using the 12 observations across the carrot products. Next, the individual regression coefficients are investigated in an exploratory manner. Another approach, we will use in the paper, is to use a mixed models, where consumers and products are treated as random effects. The `product` effect is also considered as random since we wish to consider the entire population of carrot products instead of only the 12 specific products investigated in this experiment. The following linear mixed effects model can then be considered:

$$y_{ijk} = b_{0j} + \beta_0 + (b_{2j} + \beta_2)\texttt{sens2}_{ij} + (b_{1j} + \beta_1)\texttt{sens1}_{ij} + c_k + \epsilon_{ijk} \tag{6}$$

where β_0, β_1 and β_2 stand for fixed intercept and two slopes, b_0, b_1 and b_2 stand for random intercept and random slopes, c stands for `product` effect. We assume the following covariance structure:

$$(b_0, b_1, b_2) \sim N(0, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}),\ c \sim N(0, \sigma_c^2),\ \epsilon_{ijk} \sim N(0, \sigma^2)$$

5. Types of hypotheses tests

Types I, II and III ANOVA tables as defined in SAS software SAS (1978) are provided by the **lmerTest** package. The Type I performs the sequential decomposition of the contributions of the fixed-effects and is the one produced by the `anova` method of the **lme4** package. The Type I is order dependent compared to the Types II and III, which do not depend on the order the effects are entered in the model. In terms of the hypotheses tests, the three types are the same in balanced cases, where number of observations (experimental units) at each

factor-level combination are equal. For illustration, let us consider the `TVbo` data and the model for response variable `Sharpnessofmovement` in Equation (5).

Since the `TVbo` data is balanced all the types produce the same tests. Following Searle (1987) the hypothesis test for the interaction effect γ is the following one:

$$\gamma_{i'j'} - \gamma_{ij'} - \gamma_{i'j} + \gamma_{ij} = 0 \qquad \forall i, i', j, j' \tag{7}$$

The hypothesis test for the main α effect is the following one:

$$\alpha_i - \alpha_{i'} + (1/4) \sum_j (\gamma_{ij} - \gamma_{i'j}) = 0 \qquad \forall i, i' \tag{8}$$

which is easy to interpret, that is the test for the effect of `TVset` factor averaged over all levels of the `Picture` factor is performed. In the unbalanced cases the tests for the higher order terms are still the same, whereas for the lower-order terms the hypotheses differ between the types. For example, if for some reason some observations were missing in the `TVbo` data, the Types I and II for the main α effect would no longer produce the test from Equation (8). In unbalanced situations the Types I and II hypotheses become dependent on the number of observations (experimental units) at each factor-level combination, so the hypotheses for these types become hard to interprate (Searle 1987). On the contrary, the Type III hypothesis test is the same whether the data is balanced or not, so the test for the α effect would still be the one from Equation (8).

There have been many debates regarding which type of ANOVA table is the most appropriate and when. We do not touch this topic here and refer to Speed, Hocking, and Hackney (1978); Senn (2007); Langsrud (2003); Macnaughton (2009) for the discussions. In the **lmerTest** package instead we provide a tool for obtaining the three types of ANOVA tables for the `lmer` objects, which are implemented via calculation of the appropriate hypothesis contrast matrix L in Equation (3). The calculation of the Type III L contrast matrix is based on the approach proposed by Goodnight (1978) for `proc glm` procedure in SAS software. The algorithms for constructing the Types I - III L contrast matrices are given in Appendix B.

6. Least square means and differences of least square means

The least squares means (also called population means) were introduced by Harvey (1975). The least squares means are estimates of the class or subclass means that would be expected if there would have been equal subclass numbers. For illustration let us again consider the `TVbo` data and the model for response variable `Sharpnessofmovement` in Equation (5).

The expectation, for instance, for level i of `TVset` effect is:

$$E(\overline{y}_{i\cdot}) = \mu + \alpha_i + 1/4 \sum_j (\beta_j + \gamma_{ij}) \tag{9}$$

The `TVbo` data is balanced, so the expectation is estimated by the corresponding mean: $\overline{y}_{i\cdot}$. In an unbalanced case, like e.g. if some observations were missing from the data, the expectation is no longer estimated by the corresponding mean and Equation (9) is no longer valid. The least square means are then defined in a way that Equation (9) still holds even for unbalanced data.

Generally one is interested in testing the significance about the differences of least square means. In a linear mixed effects model specified in the following form: $E(Y) = X\beta$ the null hypothesis of equality of difference of least squares means is

$$H_0 : l\beta = 0 \tag{10}$$

where l is a contrast vector. For instance, from Equation (9) the null hypothesis of equality of levels 1 and 2 for `TVset` factor is $H_0 : \alpha_1 - \alpha_2 + (1/4)\sum_j (\gamma_{1j} - \gamma_{2j}) = 0$. The t statistic for the hypothesis in Equation (10) is then:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{11}$$

where $\hat{C}$ is an estimated variance-covariance matrix of $\hat{\beta}$. Generally the t statistic does not follow a t distribution. Giesbrecht and Burns (1985) proposed a method for determining a t-distribution that approximates the distribution of t under the null hypothesis based on Satterthwaite's method-of-moment approximation to the degrees of freedom. We have implemented their work, the algorithm is in Appendix A. The confidence intervals are then computed using the following formula:

$$CI = l\hat{\beta} \pm t_{\frac{\alpha}{2}}(\nu) * \sqrt{l\hat{C}l^\top}$$

where ν is calculated using the Satterthwaite's method of approximation.

The `lsmeans` and `difflsmeans` functions from the **lmerTest** produce least square means and differences of least square means accordingly with 95% confidence intervals for all factors, that are part of an `lmer` object. The construction of l vectors for the least square means uses the `popMatrix` function from the **doBy** package (Højsgaard, with contributions from Jim Robison-Cox, Wright, Leidi, and others. 2014). The l vectors for differences of least square means are then constructed as pairwise differences of ls vectors from the least square means.

7. Step-down model-building approach

A practical data-driven approach suggested in Zuur, Ieno, Walker, Saveliev, and Smith (2009); Diggle (2002) is a step-down strategy. The strategy is based on construction of a maximal possible model followed by deletion of effects with high p-values obeying the principle of marginality. In the **lmerTest** package we have implemented a `step` function that automates the step-down approach. An outline of the algorithm is given here:

Step 1: Simplification of the random-effects structure

1. let M be the linear mixed effects model specified by a user
2. if there are random-effects in M then go to 3, otherwise stop
3. for each random-effect r_i in M do:
   (a) create a reduced model M_i by eliminating r_i from M
   (b) calculate p_i, the p-value from the likelihood ratio test of comparing M to M_i.
   (c) save p_i and M_i

4. find p_{max}; the maximum of all p_i and let M_{max} denote the corresponding model
5. set M to M_{max}. If p_{max} is higher than α level then go back to 3, otherwise stop.

If the initial model is a random-coefficient model, then the principle of simplification of the random effects is similar - the effect that contains slopes and intercept and correlation between them is incrementally reduced by removing first non-significant slopes and then non-significant intercepts. So when the effect is eliminated then the relevant correlations are eliminated as well. In Appendix C an example illustrating the process of the simplification of an error structure in random coefficient models is given.

Step 2: Simplification of the fixed-effects structure

1. consider M, the output model from **Step 1**
2. Construct an ANOVA table for M, calculate F-statistics and p-values for each fixed-effects term.
3. consider the highest order interaction effects in M. The effect with the highest p-value (p_{eff}) is identified and a model without this effect M_{eff} is constructed
4. set M_{eff} to M. If p_{eff} is less than α level or if there are no more fixed-effects then stop, otherwise go to 2

Model M from **Step 3** is the final model selected by the algorithm.

The `step` method of the **lmerTest** package contains arguments that make the step-down approach flexible. For instance, by setting the argument `reduce.random` to `FALSE` the **Step 1** can be omitted. Similarly, by setting the argument `reduce.fixed` to `FALSE` the **Step 2** can be omitted. One may specify which effects should be part of the model anyways by specifying the names of the terms in `keep.effs` argument. For example, in the `TVbo` data it may be natural to retain `Assessor` effect in the model even if the effect is not significant. By default the α level in tests for the fixed effects is 0.05 and the α level in tests for the random effects is 0.1. However both α levels can be easily changed.

8. Application of the methods

8.1. The merModLmerTest class

In the **lmerTest** package we specify a new class with the name `merModLmerTest`, which contains the `lmerMod` class from the **lme4** package:

```
R> merModLmerTest <- setClass("merModLmerTest", contains = c("merMod", "lmerMod"))
```

So if the **lmerTest** package is loaded, then the models specified with the `lmer` function are coming from the `merModLmerTest` class and not `lmerMod`. Then we define the `summary` and `anova` methods for the `merModLmerTest`, which are the extensions of the `summary` and `anova` methods of the `lmerMod` class. The nice feature about the `merModLmerTest` class is that all the methods provided by the **lme4** package for the `lmer` objects are also available for the `merModLmerTest` class. This means that by loading the **lmerTest** package and by specifying

model with the `lmer` method the users of the **lme4** package get all the methods, provided by the **lme4** package plus extended ones such as `summary` and `anova` methods and additional ones such as `step`, `lsmeans` and `difflsmeans`.

8.2. The anova method for lmer objects

Let us now consider the `TVbo` data.

`lmer` call to model in Equation (5) is:

```
R> tv <- lmer(Sharpnessofmovement ~ TVset*Picture+
+  (1|Assessor) +(1|Assessor:TVset) + (1|Assessor:Picture), data=TVbo)
```

With the following call we obtain an ANOVA table that comes from the **lme4** package

```
R> anova(tv)

Analysis of Variance Table
             Df Sum Sq Mean Sq F value
TVset         2  1.765  0.8825  0.2437
Picture       3 51.857 17.2857  4.7735
TVset:Picture 6 90.767 15.1279  4.1777
```

Now let us attach the **lmerTest** package and run again model `tv` and then apply the `anova` method again:

```
R> library(lmerTest)
R> tv <- lmer(Sharpnessofmovement ~ TVset*Picture+
+  (1|Assessor)+(1|Assessor:TVset) + (1|Assessor:Picture), data=TVbo)
R> anova(tv)

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
             Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
TVset         1.765  0.8825     2    14  0.2437 0.7869818
Picture      51.857 17.2857     3    21  4.7735 0.0108785 *
TVset:Picture 90.767 15.1279     6   138  4.1777 0.0006845 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We may notice that two additional columns are added with the names "DenDF" and "Pr(>F)" referring to denominator degrees of freedom and p values, which are calculated using the Satterthwaite's method of approximation. According to the p values the interaction effect is highly significant, which means that the products differ for the `Sharpnessofmovement` attribute. More than that the products differ mostly due to the `Picture` feature. We may also notice that by default the **lmerTest** package provides the Type III ANOVA table, **lme4** provides the sequential (Type I) ANOVA table. In this case all the types of hypotheses are identical since the `TVbo` data is balanced.

8.3. The summary method for lmer objects

The `summary` method for `lmer` objects in the **lmerTest** package produces an extended output of the `summary` method from **lme4** package. The extension of the output consists of degrees of freedom using Satterthwaite's (Kenward-Roger's) approximations for the t-test and corresponding p-values. To illustrate the `summary` method we consider the `carrots` data. We specify model in Equation (6) using **lme4**-syntax:

```
R> m.carrots <- lmer(Preference ~ sens1 + sens2 +
+  (1 + sens1 + sens2|Consumer) + (1|product), data = carrots)
```

Now let us look at the summary of the model:

```
R> summary(m.carrots)

Linear mixed model fit by REML t-tests use Satterthwaite
  approximations to degrees of freedom [merModLmerTest]
Formula:
Preference ~ sens1 + sens2 + (1 + sens1 + sens2 | Consumer) +
    (1 | product)
   Data: carrots

REML criterion at convergence: 3739.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6194 -0.5306  0.0190  0.6103  2.9309

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 Consumer (Intercept) 0.2095131 0.45773
          sens1       0.0002516 0.01586  -0.16
          sens2       0.0030473 0.05520   0.12  0.96
 product  (Intercept) 0.0335568 0.18319
 Residual             1.0335817 1.01665
Number of obs: 1233, groups:  Consumer, 103; product, 12

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  4.79911    0.07529 20.72100  63.740  < 2e-16 ***
sens1        0.01083    0.01503  9.16800   0.721  0.48913
sens2        0.07065    0.01728 10.94400   4.089  0.00181 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr) sens1
```

```
sens1 -0.010
sens2  0.023  0.032
```

The output is exactly as from **lme4** but with additional columns added to the fixed effects: "df" and "Pr(>|t|)". "df" refers to degrees of freedom based on Satterthwaite's approximation and "Pr(>|t|)" is the p-value for the t-test with "df" as degrees of freedom. We may also notice from the heading, that the object is of class `merModLmerTest`. We may conclude that the intercept and the slope for `sens2` are highly significant, so consumers prefer more sweet carrots. `sens1` has no significant impact on consumer preferences.

By setting argument `ddf` in the `summary` method to "Kenward-Roger" one may obtain Kenward-Roger's approximation.

The calculation using Satterthwaite approximation took around one second compared to Kenward-Roger's which took more than 1 minute. The p-values were identical up to the fourth digit for both approximations.

8.4. The step method for lmer objects

Let us consider again the `TVbo` data with the same response variable `Sharpnessofmovement`, but here we choose a different initial model than in Equation (5). Here we also include the `Repeat` effect as a random effect and consider a full model, where both random and fixed structures contain all possible main and interaction effects.

$$y_{ijklm} = \alpha_i + \beta_j + \alpha\beta_{ij} + c_k + ac_{ik} + bc_{jk} + abc_{ijk} + d_l + ad_{il} + bd_{jl} + abd_{ijl} + \epsilon_{ijklm}$$
$$c_k \sim N(0, \sigma_c^2) \text{ , } bc_{jk} \sim N(0, \sigma_{bc}^2) \text{ , } ac_{ik} \sim N(0, \sigma_{ac}^2) \text{ , } abc_{ijk} \sim N(0, \sigma_{abc}^2)$$
$$d_l \sim N(0, \sigma_d^2) \text{ , } bd_{jl} \sim N(0, \sigma_{bd}^2) \text{ , } ad_{il} \sim N(0, \sigma_{ad}^2) \text{ , } abd_{ijl} \sim N(0, \sigma_{abd}^2) \text{ and } \epsilon_{ijkl} \sim N(0, \sigma^2)$$

where α stands for the `TVset` effect, β for the `Picture` effect, c stands for the `Assessor` effect, d stands for the `Repeat` effect.

The corresponding model in `lmer` is:

```
R> tv <- lmer(Sharpnessofmovement ~ TVset*Picture +
+               (1|Assessor:TVset) + (1|Assessor:Picture) +
+               (1|Assessor:Picture:TVset) + (1|Repeat) + (1|Repeat:Picture) +
+               (1|Repeat:TVset) + (1|Repeat:TVset:Picture) +
+               (1|Assessor), data=TVbo)
```

Then we apply the `step` and save the results to an `st` variable:

```
R> st <- step(tv)
```

One may apply the `print` method on the `st` variable to view the results. Here instead we wrap the output into an `xtable` object of the **xtable** package (Dahl 2014) in order to nicely represent the results in the paper.

Table 2: Likelihood ratio tests for the random-effects and their order of elimination representing Step 1 of the automated analysis for the TVbo data for attribute Sharpnessofmovement

	χ^2	Chi.DF	elim.num	p-value
Assessor:Picture:TVset	0.00	1	1	1.00000
Repeat:Picture	0.00	1	2	1.00000
Repeat	0.00	1	3	1.00000
Repeat:TVset	0.00	1	4	1.00000
Repeat:TVset:Picture	0.00	1	5	1.00000
Assessor:TVset	2.79	1	kept	0.09491
Assessor:Picture	12.35	1	kept	< 0.001
Assessor	7.47	1	kept	0.00627

Table 3: F-tests for the fixed-effects and their order of elimination representing Step 3 of the automated analysis for the TVbo data for attribute Sharpnessofmovement

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	elim.num	Pr(>F)
TVset	1.76	0.88	2	14.00	0.24	kept	0.7870
Picture	51.86	17.29	3	21.00	4.77	kept	0.0109
TVset:Picture	90.77	15.13	6	138.00	4.18	kept	<0.001

Table 2 and Table 3 represent the **Step 1** and **Step 2** of the step-down model building approach in Section 8.4. The effects that have kept in the elim.num column are the ones that form the final reduced model given by the default type I levels ($\alpha = 0.1$ for the random effects and $\alpha = 0.05$ for the fixed effects).

From Table 2 it is seen that five random effects were eliminated. The `Repeat` effect is not part of the final reduced model. From Table 3 it is seen that the interaction effect `TVset:Picture` is significant, so the main effects are kept in the model according to the principle of marginality. We observe that indeed the simplified model is the one from Equation 5.

Least squares means and differences of least squares means tables are also part of the output from the `step` function. Here we vizualize the tables in barplots by applying the `plot` function on the `st` object. Since there are too many levels in the `TVset:Picture` effect, so the plot is hard to see, we ask to plot the barplots only for the `Picture` and `TVset` effects in the following way:

```
R> plot(st, effs = c("Picture", "TVset"))
```

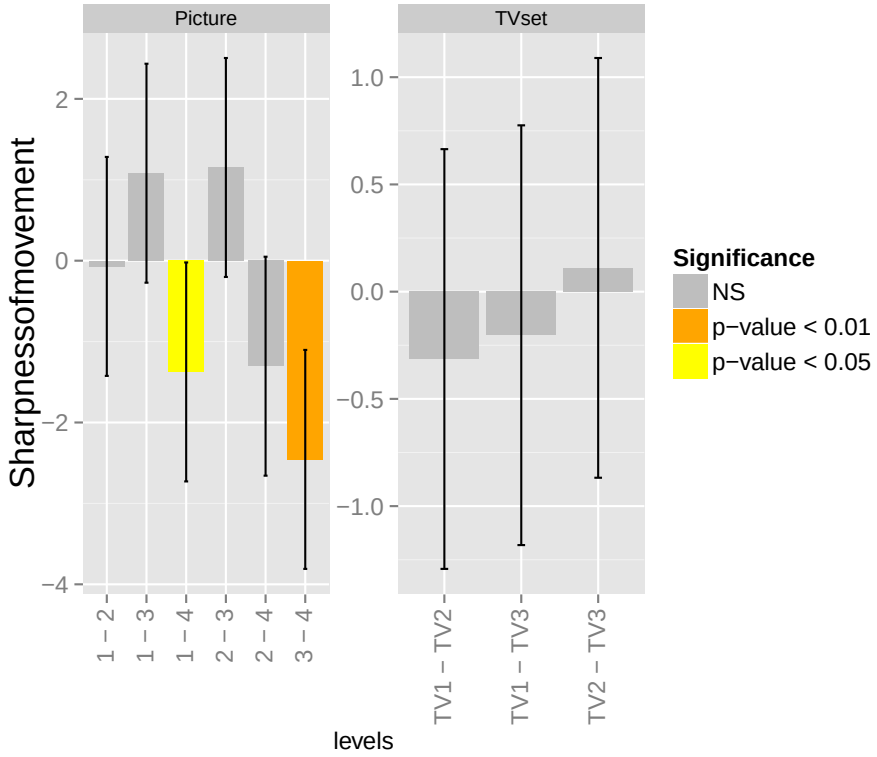


Figure 2: Barplots for differences of least square means for TVset and Picture effects together with 95% confidense intervals for the TVbo data.

The resulting plot is shown in Figure 2. The plot for the `Picture` effect shows that the most different product with respect to the `Picture` feature for the attribute `Sharpnessofmovement` is the one with level `4`. Since the `TVset` effect is non-significant according to Table 3, there are no significant differences between the levels of this effect.
There are 15 attributes in the `TVbo` data, so 14 more models should be constructed and analyzed similarily to the model for `Sharpnessofmovement` attribute considered in this example. Constructing models and applying the `step` function in a loop is therefore a useful and fast tool for getting insight into the data. More examples where the usefulness of the `step` function is illustrated are given in Kuznetsova, Christensen, Bavay, and Brockhoff (2015).

9. Computational timing issues

Halekoh and Højsgaard (2014) mention that the calculation of Kenward-Roger's approximation for some models might be computationally intensive. From our practice calculation of Satterthwaite's approximation as implementated in the **lmerTest** package requires less time than Kenward-Roger's as implemented in the **pbkrtest** package. The difference in timings depends on the size of the data and the type of the model. We have observed that for random coefficient models, the difference can be quite significant. Here we compare the computational

time for the two methods (Kenward-Roger's and Satterthwaite's) using the `carrots` data and the same model set-up as in Equation (6). In order to compare the methods for different sizes of the data, we construct 10 data sets, that are extended versions of the `carrots` data. The extension consists on replicating randomly selected rows from the `carrots` data. For example in the first data set we randomly select 1000 rows from the `carrots` data (with replacement) and then add these rows to the `carrots` data, so the size of the data becomes the size of the `carrots` data plus 1000. In the following the code for constructing the data sets and calculating the time for `anova` method applied to these data for two approximation methods is given:

```
R> size <- seq(1, 10000, by = 1000)
R> ind.size <- lapply(size,
+                     function(x) sample(seq(nrow(carrots)),
+                                        size = x,
+                                        replace = TRUE))
R> ## extend the carrots data by randomly replicating rows of the data
R> dd <- lapply(ind.size, function(x) carrots[c(1:nrow(carrots), x), ])
R> fit.mcarrots <- function(d){
+    lmer(Preference ~ sens1 + sens2
+         +(1+sens1 + sens2|Consumer) + (1|product), data=d)
+  }
R> ## apply model fit.mcarrots to all the data sets
R> m.carrots.list <- lapply(dd, fit.mcarrots)
R> ## calculate timings for the satterthwaite
R> time.sat <- lapply(m.carrots.list,
+                     function(x) system.time(anova(x))[1])
R> time.kr <- lapply(m.carrots.list,
+                    function(x) system.time(anova(x, ddf = "kenw"))[1])
```

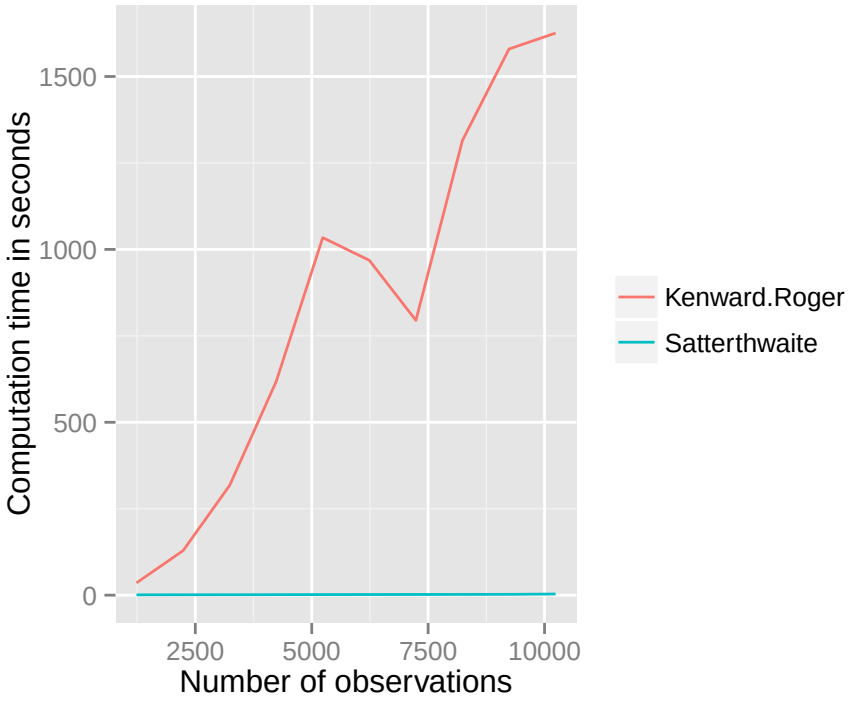


Figure 3: Differences in computational time between Kenward-Roger's and Satterthwaite's approximations for random coefficient model. carrots data

Figure 3 shows the differences in computational time between the methods. So for example for data with around 10000 observations Kenward-Roger's method took more than 1500 seconds (around 25 minutes) compared to Satterthwaite's that took around few seconds. The comparisons in computational time were made using the following hardware configuration: processor Intel(R) Core(TM) i5-3320M 2.60GHz with 2 cores (4 threads) and 8 GB of memory. We should emphasize that the comparisons were made with the **0.4-2** version of the **pbkrtest** package, since the authors of the **pbkrtest** package mention that some improvements in computational time might be in the future versions.

10. Discussion and Conclusion

In this paper we have presented our implementation of Satterthwaite's method of approximation to one and multi - degree of freedom tests. The Kenward-Roger's approximation, which is implemented in the **pbkrtest** is also available as an option in the **lmerTest** package. Then it is up to the user to decide which approximation to use or whether to use at all. From our practice, the p values that the approximation methods provide are generally very close to each other. Schaalje, McBride, and Fellingham (2002) performed a number of simulations in order to investigate the appropriateness of the approximation methods. They discovered that complexity of the covariance structures, sample size and imbalance affect the performance of both approximations. However these factors affect Satterthwaite's method more

than Kenward-Roger's. Still we believe that the Satterthwaite's method can be considered as a good alternative as it outperforms LRT in cases with unbalanced and/or small sample designs, generally is faster than Kenward-Roger's method and sometimes quite significantly. The reason that the LRT is so widely used is also connected with the fact that it is very easy and fast to apply it - just use the `anova` method to two nested models. To maintain the user-friendlyness we have wrapped the approximation methods into the `anova` and `summary` methods. So now the users of the **lme4** package can get an extended version of these methods by simply attaching the **lmerTest** package.

Another contribution of the package is a generation of the Type I - III ANOVA tables. By default the Type III ANOVA table is provided be the **lmerTest**. In terms of hypotheses tests this type is the easiest one to interprate both in unbalanced cases. Nevertheless in different situations different types of ANOVA are advised (Speed *et al.* 1978; Senn 2007; Langsrud 2003; Macnaughton 2009).

We have also introduced the `step` function, which performs backward elimination of non-significant effects. In Kuznetsova *et al.* (2015) we have shown the usefullness of this tool in a number of situations in sensory and consumer studies. We do not claim it to be a tool for confirmatory analysis but rather a nice exploratory tool. Finally, we have implemented the generation of the of least square means and differences of least square means tables with Satterthwaite's approximation to degrees of freedom.

References

Bates D, Maechler M, Bolker B, Walker S (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-4, URL http://CRAN.R-project.org/package=lme4.

by Littell O, Milliken, Stroup, Wolfinger, modifications by Douglas Bates, Maechler M, Bolker B, Walker S (2014). *SASmixed: Data sets from "SAS System for Mixed Models"*. R package version 1.0-4, URL http://CRAN.R-project.org/package=SASmixed.

Dahl DB (2014). *xtable: Export tables to LaTeX or HTML*. R package version 1.7-4, URL http://CRAN.R-project.org/package=xtable.

Diggle P (2002). *Analysis of longitudinal data*. Oxford University Press. ISBN 0198524846, 9780198524847.

Fai AH, Cornelius PL (1996). "Approximate F-Tests of Multiple Degree of Freedom Hypotheses in Generalised Least Squares Analyses of Unbalanced Split-Plot Experiments." *Journal of statistical computation and simulation*, **54**, 363.

Fox J, Weisberg S (2011). *An R Companion to Applied Regression*. Second edition. Sage, Thousand Oaks CA. URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Giesbrecht F, Burns J (1985). "Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results." *BIOMETRICS*, **41**(2), 477–486.

Goodnight J (1978). "General Linear Model Procedure." *Technical report*, SAS Institute Inc.

Halekoh U, Højsgaard S (2014). "A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest." *Journal of Statistical Software*, **59**(9), 1–30. URL http://www.jstatsoft.org/v59/i09/.

Harvey WR (1975). "Least-squares analysis of data with unequal subclass numbers."

Højsgaard S, with contributions from Jim Robison-Cox UH, Wright K, Leidi AA, others (2014). *doBy: Groupwise statistics, LSmeans, linear contrasts, utilities*. R package version 4.5-13, URL http://CRAN.R-project.org/package=doBy.

Kuznetsova A, Christensen RH, Bavay C, Brockhoff PB (2015). "Automated mixed {ANOVA} modeling of sensory and consumer data." *Food Quality and Preference*, **40, Part A**(0), 31 – 38. ISSN 0950-3293. doi:http://dx.doi.org/10.1016/j.foodqual.2014.08.004. URL http://www.sciencedirect.com/science/article/pii/S0950329314001724.

Langsrud y (2003). "ANOVA for unbalanced data: Use Type II instead of Type III sums of squares." *Statistics and Computing, Stat. Comput*, **13**(2), 163–167. ISSN 09603174. doi:10.1023/a:1023260610025.

Lawless HT, Heymann H (2010). *Sensory Evaluation of Food*. Springer Science+Business Media, LLC.

Macnaughton DB (2009). "Which Sums of Squares Are Best In Unbalanced Analysis of Variance?"

Næs T, Brockhoff PB, Tomic O (2010). *Statistics for sensory and consumer science*. John Wiley and Sons Ltd.

Pinheiro JC, Bates DM (2000). *Mixed-effects models in S and S-plus*. Springer Verlag New York, LLC.

SAS (1978). "Tests of Hypotheses in Fixed-Effects Linear Models." *Technical report*, SAS Institute Inc.

Satterthwaite F (1946). "AN APPROXIMATE DISTRIBUTION OF ESTIMATES OF VARIANCE COMPONENTS." *BIOMETRICS BULLETIN*, **2**(6), 110–114. ISSN 00994987.

Schaalje GB, McBride JB, Fellingham GW (2002). "Adequacy of Approximations to Distributions of Test Statistics in Complex Mixed Linear Models." doi:10.1198/108571102726.

Searle SR (1987). *Linear models for unbalanced data*. Wiley. ISBN 0471840963, 9780471840961.

Senn S (2007). *Statistical issues in drug development electronic resource*. John Wiley and Sons.

Speed FM, Hocking RR, Hackney OP (1978). "Methods of Analysis of Linear Models with Unbalanced Data." *Journal of the American Statistical Association*, **73**(361), 105. ISSN 01621459, 1537274x.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009). *Mixed effects models and extensions in ecology with R*. Springer Science+Business Media, LLC.

Appendices

A. F and t - statistics and the Satterthwaite's approximation

Assume we have the mixed model in Equation ((1))

with X the $n \times p$ design matrix for the fixed-effects and Z the $n \times k$ design matrix for the random-effects.

The variance of y is therefore

$$V(\theta) = ZG(\theta)Z^\top + R(\theta)$$

Where parameter θ consist's of residual error variance and variance of random-effects.

The variance-covariance matrix of β is

$$C(\theta) = (X^\top V(\theta)^{-1} X)^{-1} = (X^\top (ZG(\theta)Z^\top + R(\theta))^{-1} X)^{-1}$$

For simplicity further we will suppress θ in the notation.

Giesbrecht and Burns (1985) investigated a one-degree test of hypothesis $H_0 = l^\top \beta$ where l is a vector. A corresponding t - statistic is then:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{12}$$

where $\hat{C} = C(\hat{\theta})$

They followed Satterthwaite (1946) and assumed that the quantity

$$\frac{df(l^\top \hat{C} l)}{(l^\top C(\theta) l)}$$

approximately follows chi-square distribution. Then they used Satterthwaite's method-of moments approximation to degrees of freedom:

$$df = \frac{2(l^\top \hat{C} l)^2}{[var(l^\top \hat{C} l)]}$$

Taking $f(\theta) = l^\top C(\theta) l$, $var(f(\theta))$ can be approximated by applying univariate delta method as:

$$var(f(\theta)) \approx [\nabla_{f(\theta)} \hat{\theta}]^\top A [\nabla_{f(\theta)} \hat{\theta}]$$

where $\nabla_{f(\theta)} \hat{\theta}$ is a vector of partial derivatives of $f(\theta)$ with respect to θ evaluated at $\hat{\theta}$. A is the variance covariance matrix of the $\hat{\theta}$-vector, which can be found from the second derivatives of the log-likelihood function. Matrix A is not directly extractable from the **lme4** package. In the **lmerTest** package we specify the deviance function with respect to θ parameters and find the second derivatives on the optima $\hat{\theta}$. Similarly we specify a function that calculates variance-covariance with respect to the θ parameters. Then we calculate partial derivatives evaluated at the optima.

In a multi-degree of freedom test a hypothesis of interest is $H_0 : L\beta = 0$, where L is an estimable contrast matrix of $q = rank(L) > 1$. A commonly used test statistic for this hypothesis is:

$$F = \frac{(L\hat{\beta})^\top (L\hat{C}L^\top)^{-1} (L\hat{\beta})}{q} \tag{13}$$

Even though the statistic is called F, it usually does not follow an F distribution. Fai and Cornelius (1996) proposed a method for approximating distributions of F. There they also used the Satterthwaite's method-of-moment approximation to the degrees of freedom. First they decomposed $(L\hat{C}L^\top)^{-1}$ in order to yield $P^\top (L\hat{C}L^\top)^{-1} P = D$ where P is an orthogonal matrix of eigenvectors and D is a diagonal matrix of eigenvalues. Using this decomposition, $Q = qF$ can be written as a sum of q squared t variables,

$$Q = \sum_{m=1}^{q} \frac{(PL\hat{\beta})_m^2}{D_m} = \sum_{m=1}^{q} t_{\nu_m}^2$$

where $(PL\hat{\beta})_m$ denotes the mth element of $PL\hat{\beta}$ and D_m is the mth diagonal element of D. Then Fai and Cornelius noted that each ν_m can be approximated by the Giesbrecht-Burns single degree-of-freedom method:

$$\nu_m = \frac{2D_m}{g_m^\top A g_m}$$

where g_m is the gradient of $l_m C l_m^\top$ with respect to θ with l_m being the mth row of PL. Using the relationship $E(F_{q,\nu}) = \frac{\nu}{\nu-2}$ for $\nu > 2$, they then find ν such that $q^{-1}Q \sim F_{q,\nu}$ approximately. Since the t_{ν_m} can be regarded as having independent Student's t - distributions with ν_m degrees of freedom,

$$E(Q) = \sum_{m=1}^{q} E(t_{\nu_m}^2) = \sum_{m=1}^{q} E(F_{q,\nu}) = \sum_{m=1}^{q} \frac{\nu_m}{\nu_m - 2} = E_Q$$

Now since

$$\frac{1}{q}E_Q = \frac{\nu}{\nu - 2}$$

it can be shown that

$$\nu = \frac{2E_Q}{E_Q - q}$$

B. Hypotheses contrast matrices

The key step in constructing the F-test for an effect is in constructing the constrast matrix defining the hypothesis appropriately.

B.1. Notations and definitions

complete rank deficient matrix

let X be a design matrix. It can be partitioned according to the model effects:

$$X = [1|X_2|...|X_p] \tag{14}$$

The design matrix is usually assumed to have full (column) rank. If (some of) the model effects are factors, then the matrix will not be of full rank (it will be turned into full rank matrix by deletion of a selection of columns).
Let X^+ denote the design matrix before reduction of full column rank. This matrix is generated in **lmerTest** by generating the design matrix for each effect separately and then concatenating as in 14.

Estimable functions

Function of parameters that are unaffected by the choice of model parametrization are called *estimable functions of the parameters.* A linear function of the parameters $L\beta$ is estimable if and only if L is in the row-space of X (put reference here). Therefore rows of X form a

generating set from which any estimable L can be constructed. Since the row spaces of X, X^tX, $(X^tX)^-(X^tX)$ are identical, they all form generating sets for any estimable L (SAS Users guide ch. 12). $(X^tX)^-(X^tX)$ has the property of containing lots of zeros, so it is used as a *basis set of estimable functions*:

$$L = (X^tX)^-(X^tX) \tag{15}$$

. Here $^-$ is understood as a generalized inverse, X is as the complete (rank-deficient) design matrix (we denoted it previously X^+).

Contained effects

Consider two effects: e_1 and e_2. Then e_1 is said to be contained in e_2 if

1. all factors associated with e_1 (if any) are also associated with e_2
2. there are more factors with e_2 than with e_1
3. both effects involve the same continious variables (if any)

NOTE: Consider the intercept (μ) as having no continuous variables and no classes.

B.2. Type III hypotheses contrast matrices

Here we refer to the rules of generating Type 3 hypotheses matrices, as proposed in Goodnight (1978). Let L be the *general set of estimable function* (15), e be an effect, for which we want to construct hypothesis matrix, say L^e. Then the following rules create hypothesis matrix L^e:

Rule 1 using row operations, the rows in L not related to e are set to zero

1. Find all columns of L not related to e: $j = 1, ..., J$
2. For each j, find all non-zero elements in $L[, j]$; $i = 1, ..., n_j$
3. For each $i \in n_j$: $L[i,] \longleftarrow L[i,]/L[i, j]$
4. Set the ith row to zero: $L[i,] \longleftarrow 0$

Rule 2 Find a basis set of estimable functions for e. This rule is needed only when other effects contain e and are entered in a non-standard order. In the **lmerTest** the model is updated at the beginning, so the effects are entered in a standard order. Hence **Rule 2** is skipped.

Rule 3 Effects that *contain* e are orthogonalized to e in the basis set for e. Starting with the first row in L having all zeros associated with e all other rows are made orthogonal to it using row operations, the row is then set to zero. This is done for all other rows having all zeros associated with the e.

B.3. Type I hypotheses contrast matrices

Following SAS (1978) the Type I hypothesis contrast matrix L is the Forward-Dolittle transformation of the X^tX with each nonzero row divided by its diagonal, where X is a rank

deficient design matrix in Equation (14). Then the contrast matrix L^e for an effect in question e is the corresponding to the effect e rows of the L matrix.

B.4. Type II hypotheses contrast matrices

Following SAS (1978) the Type II hypothesis contrast matrix L^e for an effect in question e is calculated in the following way:

1. the columns of the design matrix X in Equation (14) are rearranged in a way that columns corresponding to effects that do not contain the effect e are put before the columns corresponding to the effect e. Let us denote this rearranged design matrix X'
2. the L matrix is calculated as the Forward-Dolittle transformation of the ${X'}^t X'$ with each nonzero row divided by its diagonal
3. the columns of L are rearranged to reflect the original order of the model
4. the contrast matrix L^e is the corresponding to the effect e rows of the L matrix

C. Example of analysis of the error structure in a random coefficient model

Let us consider a model in Equation (6). The error structure of this model is:

$$(b_0, b_1, b_2) \sim N(0, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}), \quad c \sim N(0, \sigma_c^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2) \tag{16}$$

Let us specify it via the `lmer` function:

```
R> m.carrots <- lmer(Preference ~ sens1 + sens2 +
+  (1 + sens1 + sens2|Consumer) + (1|product), data = carrots)
```

Then we apply a `step` function from the **lmerTest** package, requiring not to perform tests on the fixed effects since we are not interested in them in this example:

```
R> step(m.carrots, fixed.calc = FALSE)

Random effects:
               Chi.sq Chi.DF elim.num p.value
sens1:Consumer   1.83      3        1  0.6090
sens2:Consumer   7.81      2        0  0.0202
product         16.16      1        0   1e-04

Final model:
lme4::lmer(formula = Preference ~ sens1 + sens2 + (1 | product) +
    (sens2 | Consumer), data = carrots, REML = reml.lmerTest.private,
    contrasts = l.lmerTest.private.contrast, devFunOnly = devFunOnly.lmerTest.private)
```

The first row in the random effects table means that the LRT was applied to model `m.carrots` and a reduced one, that does not contain random slope `sens1`. We can see that in the following code:

```
R> m.carrots.red.sens1 <- lmer(Preference ~ sens1 + sens2 +
+  (1 + sens2|Consumer) + (1|product), data = carrots)
R> anova(m.carrots, m.carrots.red.sens1, refit = FALSE)

Data: carrots
Models:
..1: Preference ~ sens1 + sens2 + (1 + sens2 | Consumer) + (1 | product)
object: Preference ~ sens1 + sens2 + (1 + sens1 + sens2 | Consumer) +
object:     (1 | product)
       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
..1     8 3757.3 3798.2 -1870.7   3741.3
object 11 3761.5 3817.8 -1869.7   3739.5 1.8274      3      0.609
```

The degrees of freedom in this test are equal to 3, meaning that the tests were made for three parameters: random slope for `sens1` (σ_1^2) and correlations between the random slope `sens1` and the random slope `sens2` (σ_{12}) and the intercept (σ_{01}). Model `m.carrots.red.sens1` is the final reduced model (the "elimn.num" column is equal to 0 for the rest of the rows in the random effects table meaning that the random slope `sens2` and the intercept are kept in the model according to the default Type 1 error equal to 0.1). The error structure of the final reduced model is then:

$$(b_0, b_1, b_2) \sim N(0, \begin{pmatrix} \sigma_0^2 & \sigma_{02} \\ \sigma_{02} & \sigma_2^2 \end{pmatrix}), \quad c \sim N(0, \sigma_c^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Affiliation:
Alexandra Kuznetsova
Department of Applied Mathematics and Computer Science
Statistics and Data Analysis Section
DTU Compute
Richard Petersens Plads
Building 324, DK-2800 Kgs. Lyngby
E-mail: alku@dtu.dk

Appendix C

Analysing sensory data in a mixed effects model framework using the R package SensMixed

Kuznetsova A., Amorimb I., Brockhoff P. B., Limab R. R.

Analysing sensory data in a mixed effects model framework using the R package SensMixed

Alexandra Kuznetsova[a,*], Isabel de Sousa Amorim[b], Per Bruun Brockhoff[a]

[a] *DTU Compute, Statistical section, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark*
[b] *DEX - Departamento de Ciências Exatas, Universidade Federal de Lavras, Campus da UFLA - Caixa Postal 3037 Lavras, MG, Brasil*

Abstract

In this paper we present the open-source R package SensMixed Kuznetsova et al. (2013c), that is dedicated to analyze sensory data within a mixed effects model framework. The package offers mixed model ANOVA specifically prepared for multi attribute sensory data adopting the generality of the lmerTest package Kuznetsova et al. (2013a) like e.g. allowing for multi-way product structures, incomplete data, complex error structures and providing the new tool introduced in Brockhoff P. B. (2015) to visualise the results. Also in Brockhoff et al. (2015) a new mixed model approach was suggested, the Mixed Assessor Model (MAM) that properly takes into account possible scale range differences. However there the methodology considers only a rather simple 2-way setting. In this paper we consider an extended version of the MAM and the SensMixed package provides a tool to construct and visualize the results of analysis of such models. Finally the SensMixed package provides an intuitive and easy-to-use graphical user-interface that handles all statistical computations in the background and visualises results in different types of plots and tables. The usefulness of the presented tools is illustrated on two sensory studies.

Keywords: sensory profiling, mixed models, multi-way product structure, **R** program

1. Introduction

Mixed effects models form an integral part of statistical analysis of sensory as well as consumer data. In sensory studies the analysis of variance

*Corresponding author. E-mail address: alku@dtu.dk (A. Kuznetsova).

techniques (ANOVA) are generally applied in order to extract important attribute-wise product difference information (Lawless & Heymann, 2010). Typically the assessor-by-product interaction forms the error structure in significance tests for product effects (Lawless & Heymann, 2010). Still this approach cannot handle all possible situations and as pointed out in Kuznetsova et al. (2015), it is important generally, and specifically within the sensory and consumer field to be able to also handle more complicated settings. In Kuznetsova et al. (2015) an automated analysis of more complex settings together with the R package named lmerTest that performs the analysis was introduced. An example showing the usefulness of the package applied to sensory data was also presented there. This was a one step forward in facilitating analysis of sensory data in complex situations such as multi-way product structures, unbalanced data and complex error structures. In this paper we present a number of tools to analyze and visualize the results of analysis of sensory data within a mixed effects model framework. The first one uses the same technique of the automated analysis as in Kuznetsova et al. (2015) but applied simultaneously to all attributes and presenting the results in a compact and efficient way.

The so-called mixed assessor model (MAM) was proposed in Brockhoff et al. (2015), that corrects for a possible scaling effect. There the authors showed that considering such models is more appropriate for sensory data whenever the scaling effect is present. One of the main advantages of the MAM is that they produce improved hypothesis tests for product effects. However a rather simple 2-way MAM together with the tool for analyzing it was proposed there. In this paper we propose an extended versions of MAM where a possible multi-way product structure can be accounted for together with the 3-way error structure, where a replicate effect is also accounted for. The SensMixed provides as well the tools to analyze the extended versions of MAM in providing the tests for the fixed effects as well as random effects together with the post-hoc product pairwise comparisons.

The visual tool for analysis of product effects based on effect size measures introduced in Brockhoff P. B. (2015) is also implemented in SensMixed. The idea presented in Brockhoff P. B. (2015) is to interpret effects relative to the residual error and to choose the proper effect size measure. It has been shown that the effect size estimate presentend in Brockhoff P. B. (2015), so-called d-tilde, has a close link to the Thurstonian d-prime, and as such is a generic measure that can be interpreted and compared across any attribute and situations.

All presented techniques are graphically oriented and should therefore be easy to understand by sensory practitioners and non-statisticians. This allows for efficient analysis of sensory data, enables the practitioner and non-statistician to focus on results of the statistical analysis rather than spending time on trying to apply algorithms on the data by themselves.

2. Data

2.1. Sensory data of cherry products

The data comes from (put reference), and is an example of a multi-way product structure sensory data. The purpose of this study was to assess nine fruit drinks, which were made from three levels of lime flavour and three levels of fibre (see Table 2). So all in all 9 products (3-by-3 combinations of flavour and fiber) were assessed by 10 trained panelists in 3 replications for the following 13 attributes: Cherry aroma, Apple aroma, Lime zest aroma, Unfresh aroma, Metallic aroma, Cherry flavour, Apple flavour, Lime zest flavour, Sweet taste, Sour taste, After taste, Astringency, Creamy.

[Table 1 about here.]

2.2. Sensory study of car audio systems

The data comes from the company Bang and Olufsen A/S, Struer, Denmark. The purpose of this study was to rate products, specified by three features: Car (sound system), SPL (reproduction of sound pressure level) and Track (music program). The trained audio panel was composed by 10 assessors (Participant) who evaluate 90 products (CLIP) for 8 different response variables (Attributes) in 2 replications. Only 8 assessors completed both replications.

The names of the 8 attributes were transtaled from Danish [1] as: continuous noise, accuracy in the lower frequency range, accuracy in the upper frequency range, reverberation, stereo effect, strength of the bass range, strength of the treble range, strength of the mid-range. By simplicity we call them att1, att2, ..., att8.

[1] Kontinuerligstøj, Præcisioninedreområde, Præcisionivreområde, Rumklang, Stereovirkning, Styrkenafbas, Styrkenafdiskant og Styrkenafmellemtone

3. Theory

3.1. multi-way product structure data in sensory studies

Sometimes in sensory as well as consumer studies products are formed by a combination of features (Jaeger et al., 2013; Beck et al., 2014). Both data sets considered here are examples of such studies. For instance, in the cherry data 9 products are formed by 3-by-3 combinations of Flavour and Fiber features. An approach to analyze such data could be considering one product factor with 9 levels and considering a 2-way ANOVA mixed model and the assessor-by-product interaction as the error structure (Lawless & Heymann, 2010). This approach can be easily done via, for instance, the PanelCheck software (Nofima Mat, 2008) .

Another approach, where the multi-way product structure can be accounted for is captured in the following model:

$$\begin{aligned} y_{ijkl} = \mu &+ \texttt{fi}_j + \texttt{fl}_k + \texttt{fifl}_{jk} \\ &+ a_i + a\texttt{fi}_{ij} + a\texttt{fl}_{ik} + a\texttt{fifl}_{ijk} + \epsilon_{ijkl} \end{aligned} \tag{1}$$

$$\begin{aligned} a_i &\sim N(0, \sigma^2_{assessor}) \\ a\texttt{fi}_{ij} &\sim N(0, \sigma^2_{assessor \times fiber}) \\ a\texttt{fl}_{ik} &\sim N(0, \sigma^2_{assessor \times flavour}) \\ a\texttt{fifl}_{ijk} &\sim N(0, \sigma^2_{assessor \times flavour \times fiber}) \\ \epsilon_{ijkl} &\sim N(0, \sigma^2_{error}) \end{aligned}$$

where now there are two main effects `fi` and `fl` corresponding to factors Fiber and Flavour and an interaction effect `fifl` corresponding to interaction effect between Flavour and Fiber. Kuznetsova et al. (2015) showed that accounting for the multi-way product structure gives more insight into the data.

3.2. Complex error structures in sensory studies

In the cherry data the assessors scored the products in 3 replications. Hence it might be sensible to test the replicate effect as well. An extension to model 1 can be considered, where a replicate effect and its interaction with the other effects are additionally added, which results in construction of the following linear mixed effects model:

$$\begin{aligned} y_{ijkl} = \mu &+ \texttt{fi}_j + \texttt{fl}_k + \texttt{fi}_j\texttt{fl}_k \\ &+ a_i + a\texttt{fi}_{ij} + a\texttt{fl}_{ik} + a\texttt{fifl}_{ijk} \\ &+ r_l + ar_{il} + \texttt{fi}r_{jl} + \texttt{fl}r_{kl} + \texttt{fifl}r_{jkl} + \epsilon_{ijklr} \end{aligned} \tag{2}$$

$$
\begin{aligned}
a_i &\sim N(0, \sigma^2_{assessor}) \\
a\texttt{fi}_{ij} &\sim N(0, \sigma^2_{assessor \times fiber}) \\
a\texttt{fi}_{ik} &\sim N(0, \sigma^2_{assessor \times flavour}) \\
a\texttt{fifl}_{ijk} &\sim N(0, \sigma^2_{assessor \times flavour \times fiber}) \\
r_l &\sim N(0, \sigma^2_{replicate}) \\
ar_{il} &\sim N(0, \sigma^2_{assessor \times replicate}) \\
\texttt{fi}r_{jl} &\sim N(0, \sigma^2_{fiber \times replicate}) \\
\texttt{fl}r_{kl} &\sim N(0, \sigma^2_{flavour \times replicate}) \\
\texttt{fifl}r_{jkl} &\sim N(0, \sigma^2_{fiber \times flavour \times replicate}) \\
\epsilon_{ijkrl} &\sim N(0, \sigma^2_{error})
\end{aligned}
\tag{3}
$$

From Equations 3 we observe that 9 random effects form the random part of the extended model. It might be that not all of these effects contribute to the systematic variation in the data and therefore could be excluded from the model (Kuznetsova et al., 2015). In the SensMixed the `step` method from the lmerTest package is used, that finds a parsimonious random structure by sequentially removing non-significant random effects.

3.3. Mixed Assessor Model (MAM)

There is a large literature on studying and monitoring individual differences in sensory profile data. The difference in use of the scale is generally considered to be part of nuisance effects (Næs, 1990; Tomic et al., 2013), which may be either reduced by extended training or accounted for in data analysis. In Brockhoff et al. (2015) the new mixed model approach was suggested, the Mixed Assessor Model (MAM) that properly takes into account scale range differences.

The Mixed Assessor Model (MAM) can be specified in the following form:

$$
y_{ijk} = \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + \varepsilon_{ijk} \tag{4}
$$

$$
a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)
$$

where a_i is the assessor main effect, $i = 1, 2, ..., I$, the ν_j the product main effect, $j = 1, 2, ..., J$, $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and hence β_i is the individual (scaling) slope (with $\sum_{i=1}^{I} \beta_i =$

0), the d_{ij} is the random interaction term, that captures the disagreements between the assessors. (Brockhoff et al., 2015) showed that MAM produces valid and improved hypothesis tests for as well overall product differences as post hoc product difference testing.

3.4. Extended MAM

The MAM in Equation 4 considers a rather simple 2-way structure. As pointed out in Section 3 it is sensible to consider more complex structures such as 3-way, where the replicate/session effect forms also part of the model as well as multi-way product structures. All that calls for a need in considering extended versions of MAM, where scaling effect can be part of a more complicated linear mixed effects model.

3.4.1. 3-way MAM

The 3-way linear mixed assessor model can be specified in the following form:

$$y_{ijkl} = \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + r_k + ar_{ik} + a\nu_{jk} + \varepsilon_{ijkl} \tag{5}$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ r_k \sim N(0, \sigma^2_{replicate}),$$
$$ar_{ik} \sim N(0, \sigma^2_{assessor\times replicate}),\ \nu r_{jk} \sim N(0, \sigma^2_{product\times replicate}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

from which we may notice that three more random effects (replication effect and interactions between replication and the other effects) form part of the random structure compared to MAM in Equation 4. ν_j is again an effect corresponding to the product factor.

3.4.2. Multi-way product structure MAM

It is not uncommon that the products investigated are formed as combinations of features. The multi-way product structure version of the MAM is constructed in the same way as in Section 3.1, that is product factor ν_j is replaced by the main feature effects and interactions between them. So, for instance, for the cherry data the multy-way product structure MAM, combined with the 3-way error structure can be specified in the following form:

$$y_{ijkl} = \mu + a_i + \texttt{fi}_j + \texttt{fl}_k + \texttt{fifl}_{jk} + \beta_i x_j + d_{ij} + r_k + ar_{ik} + a\nu_{jk} + \varepsilon_{ijkl} \tag{6}$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ r_k \sim N(0, \sigma^2_{replicate}),$$
$$ar_{ik} \sim N(0, \sigma^2_{assessor\times replicate}),\ \nu r_{jk} \sim N(0, \sigma^2_{product\times replicate}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

We may observe that the random part does not account for the multi-way product structure - this is done in purpose

3.4.3. Post-hoc in MAM

the product averages of MAM model in Equation 4 are 100% confounded with the product effect ν_j . As also pointed out in Brockhoff et al. (2015), due to this confounding the product contrasts are not easily obtained. We used an approach of calculating the product contrasts using the version of model without the scaling effect, but in tests using the variance-covariance matrix $\hat{C}$ coming from MAM. The details of the approach are given in Appendix 7.1

4. Methods

The SensMixed package includes an application that has a graphical user interface (GUI) and that is implemented via the R package named shiny (Chang et al., 2015). Apart from providing the GUI for advanced statistical methods within a mixed effects framework, the application includes such crucial functionalities as importing the data in different formats, presenting results in tables and plots as well as saving them. In order to run the application, one needs to install the SensMixed package and call the `SensMixedUI` function by typing in the R console the following lines:

```
library(SensMixed)
SensMixedUI()
```

In the following section the methods that form the SensMixed package will be discussed. The results of these methods are visualized in various plots and tables helping sensory practitioners to visually detect performance issues without having to know all details on the statistical methods.

4.1. SensMixed modeling controls

A number of options for the mixed effects model building is introduced in the SensMixed package, which make the model building more flexible and advanced. The main modelling controls are the following ones:

- **error structure**

  No Rep: assessor effect and all possible interactions between assessor and product effects

  2-WAY: **No Rep**, replicate effect and interaction between assessor and replicate effects

3-WAY: assessor and replicate effect and interaction between them and interaction between them and Product effects

- **product structure**

  1 main product effects

  2 main product effects and 2-way interactions between them

  3 main product effects and all possible interactions between them

- **scaling correction**

  Yes

  No

These controls are responsible for the specification of the mixed effects model. **error structure** stands for the specification of the random part of a mixed effects model. **error structure = 3-WAY** produces the maximal possible random structure. This option is advised in (Kuznetsova et al., 2015). However if, for example, from the studies it is known that there is no replication effect, then the **No-Rep** option can be considered. If it is known that there is no interaction between replication and product effects, then the **2-WAY** option may be chosen, which also conducts the analysis in a faster way.

The **product structure** is responsible for specification of the fixed part of the mixed effects model. If there is no multi-way product structure in the data, then all options produce the same fixed part. Otherwise the option **3** produces the maximal possible fixed structure.

If one chooses to correct for scaling, then the MAM is constructed as in Section 3.3. According to Brockhoff et al. (2015) whenever the scaling is significant it is advisable to correct for it, since the tests for the product effects become more powerful.

According to the specified modelling controls the mixed effects models are constructed for all attributes using the lme4 package Bates et al. (2014) and then the `step` method of the lmerTest Kuznetsova et al. (2013a) is applied to each model. In all cases the fixed part is not simplified. By default the non-significant random effects are eliminated from the model according to the specified by a user Type 1 error (0.1 the default one). However one may require not to eliminate the random effects, or specify which effects should be kept in the model even if not being significant.

4.2. $\sqrt{\chi^2}$ *plots*

$\sqrt{\chi^2}$ plot represents the bars for the square root of the χ^2 statistics of the likelihood ratio test applied to random-effects for each sensory attribute. The colours of the bars represent the significance level of the effects. This plot is a valuable visualisation tool that helps the user to quickly investigate, for instance, whether there is a replication effect, or is there a disagreement between assessors on scoring the products and if yes, then according to which features. The plot is very similar to the F-plot, that the PanelCheck provides, but can account for more complex models and hence provide more information on the data. If there is a requirement for the reduction of the random effects, then the chi-squared values are the sequential ones, that is they come from the stepwise selection process based on the methodology proposed by (Kuznetsova et al., 2015).

4.3. $\sqrt{F}$ *plots for scaling effects*

$\sqrt{F}$ plots represent the bars corresponding to the values of the square root of the F-statistics in a test for a scaling effect for each sensory attribute. The colours of the bars represent the significance level. These plots are useful for detecting the scaling effects for the attributes. If, for instance, the plot shows that the scaling effect is significant, this means that the assessors use the scale differently for the attribute in question.

4.4. $\sqrt{F}$ *plots for product effects*

$\sqrt{F}$ plots represent the bars corresponding to the values of the square root of the F-statistics in a test for product effects for each sensory attribute. The colours of the bars represent the significance level. These plots are useful for detecting whether the assessors are able to discriminate the products according to the attributes. If the multi-way product structure is present, then the plot also visualizes according to which feature the products differ.

4.5. delta-tilde plots

The delta-tilde plots represent the bars for the effect size expressed in terms of relative pairwise comparisons for each product effect and each sensory attribute. The colours of the bars represent the significance level. These plots proposed by Brockhoff P. B. (2015) can be considered as a complement to the F statistics in tests for product effects. Brockhoff P. B. (2015) showed

that the effect size estimate d-tilde has a close link to the Thurstonian d-prime, and as such is a generic measure of the effect size that can be interpreted and compared across any attributes and factor levels specially when the multi-way productis present, with different number of levels and different number of observations within the levels. Brockhoff P. B. (2015) presented the estimate of delta-tilde as the back transformation of the F-statistic correcting the bias by subtracting 1:

$$\tilde{\hat{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F-1}$$

More detailed information on the statistical aspects of delta-tilde estimates are given in Brockhoff P. B. (2015), where also the algorithm for calculating them for balanced cases is proposed. SensMixed contains a generic implementation of the method, that can handle unbalanced data and complex error structures. The delta-tilde plot gives an additional insight to the data, especially in cases of multi-way product structures.

4.6. Post-hoc plots

The post-hoc plot shows bars for multiple pairwise comparison tests for products effects together with the 95% confidence intervals for each attribute. This plot is useful for detecting which products are different. When the multi-way product structure is considered, the plot is also valuable in providing information according to which features do products differ.

4.7. step output

The step output is the result of the analysis of random and fixed effects produced by the `step` function of the lmerTest package (Kuznetsova et al., 2013b). The output consists of two tables. These tables can be considered as a complement to $\sqrt{\chi^2}$ and $\sqrt{F}$ plots, as they present a more detailed information on the analysis of random and fixed effects of models.

The first table is the ANOVA-like table for the random effects, where each random effect is tested with likelihood ratio test. If a user requires reduction of the random structure, then the random effects are the sequential ones, where non-significant random effects are sequentially eliminated if being non-significant according to the specified Type 1 error rate.

The second table produces the ANOVA table for fixed effects. This table contains the F statistics together with the delta-tilde estimates and corresponding p-values in tests for product and scaling effects. The sums of squares as well as the mean squares are part of the table as well. If the reduction of the non-significant random effects is performed on the first table, then the construction of the second table uses the reduced random structure in tests for the fixed effects.

5. Results

5.1. Results for the Cherry data

5.1.1. one-way product structure

First let us consider just one product factor Sample with 9 levels (see Table 2) and choose the following modelling controls:

error structure = 3-WAY

product structure = 1

scaling correction = Yes

Figure 1 represents the sequential $\sqrt{\chi^2}$ plot. It can be seen that the replicate effect and its interaction with the other effects are non-significant for all attributes, which means that there is no systematic variation across replications.

Figure 2 shows the $\sqrt{F}$-plot for the scaling effects. From the Figure it is clear that the scaling effect is highly significant for almost all attributes except Sour taste, Sweet taste and Astringency.

Figure 3 represents the $\sqrt{F}$-plot for the product effects. The mixed assessor models considered in this plot use the reduced random structures, as by default the non-significant random effects are eliminated in the SensMixed. Since the scaling effects are corrected, the tests for the sample effects become more powerful (Brockhoff et al., 2015). We may observe that the sample effect is highly significant for all attributes except Sweet.taste and Astringency, so for most of the attributes assessors are able to discriminate the products. In order to see a more detailed information on analysis of fixed and random effects for the attributes Sweet.taste and Astringency we take a look at the step output.

Tables 3 and 4 represent the analysis of random and fixed effects accordingly for the attribute Sweet.taste. From Table 3 we observe that assessor effect and interaction between sample and assessor are highly significant, the effect corresponding to interaction between sample and replicate was eliminated (`elim.num = 1`), as considered non-significant according to the default Type 1 error rate equal to 0.1. The significant effects with `elim.num` equal to 0 are then used in tests for the scaling and sample effects in Table 4. From Table 4 we observe that the sample effect is significant at 0.1 error rate, so we decide to keep the Sweet.taste attribute for further analysis (put reference). It can also be seen that the scaling effect is significant at 0.1 rate, and according to Brockhoff et al. (2015), it is better to keep scaling effects in the model at rate less than 0.2. The step output for the fixed effects for the Astringency attribute presented in Table 5 shows that the p-value for the sample effect is 0.516, so we decide to discard this attribute from further analysis.

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

5.1.2. multi-way product structure

As a next step we would like to account for the multi-way product structure in the mixed effects model in order to get more insight into the data. This may be achieved by considering effects Flavour and Fiber and interaction between them instead of one Sample effect in the fixed part of a mixed effects model. Therefore the following modelling controls need to be chosen:

error structure = No-Rep

product structure = 3

scaling correction = Yes

We choose **error structure = No-Rep** since we have shown that the replicate effect and its interaction with the other effects are non-significant. We do not include Astringency attribute since we have shown that the assessors are not able to discriminate the products for this attribute.

[Table 5 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

Table 6 represents the sequential chi-squared values (i.e. from the stepwise selection process) for the tests of random-effects for a multi-way product structure case. It can be observed that the Assessors disagree in scoring the products (Flavour:Fiber:Assessor effect is significant for almost all attributes), though the disagreement according to the product features is different from attribute to attribute. For instance, for the Creamy attribute it is exclusively due to the Fiber the assessors disagree in scoring the products. On the contrary, for the Lime zest aroma attribute it is mainly due to the Flavour the assessors disagree.

Figure 4 represents the d-tilde plot. From the Figure 4 we see that the interaction between Fiber and Flavour is non-significant for almost all the attributes.

We can also notice that the p value for the Fiber effect is less than 0.05 for the Sweet taste (when considering just one Sample effect we found that the p value was less than 0.1). Flavour effect is significant for all attributes related to flavour and aroma. For the unfresh aroma, lime zest flavour and lime zest aroma attributes Fiber effect is significant as well. For the unfresh aroma and lime zest aroma attributes the interaction between Fiber and Flavour effects is present. We may also notice that the Fiber effect is significant for Apple aroma attribute whereas Flavour is not according to the 0.05 Type 1 error rate. However the height of the bar corresponding to the Flavour is much higher, so the size of the Flavour effect is high. Since the delta-tilde represent the effect sizes, the sizes of the bars can be compared between each other. So, for instance, the size of the Fiber effect for the Creamy attribute is much higher than for the other attributes. The size of the Flavour effect for the unfresh aroma attribute is higher than the size of the Fiber effect.

If we want to see which levels of Fiber are different for the unfresh aroma attribute we may look at the least squares means and differences of least

squares means. From Figures 5 we may observe that with augmenting levels of flavour in products, the unfresh aroma reduces. There is a significant difference between flavour with levels Fla0 and Fla1 as well as Fla0 and Fla2, and there is no significant difference between products with levels Fla1 and Fla2 according to the unfresh aroma attribute.

5.2. Results for the Car data

5.2.1. one-way product structure

First we consider one product factor Clip with 90 levels and choose the following modelling controls:

error structure = 3-WAY

product structure = 1

scaling correction = Yes

Figure 6 represents the sequential $\sqrt{\chi^2}$ plot. It can be seen that the repetition effect and its interaction with the clip effect are non-significant for all attributes, however there is a significant interaction between participants and repetition. We can as well observe that Clip:Participant is significant for all attributes. Since here the mixed assessor models are considered, the Clip:Participant effect means the real disagreement between participants in scoring the products.

Figure 7 shows the $\sqrt{F}$-plot for the scaling effects. From the Figure it is clear that the scaling effect is significant for all attributes, so the participants use the scale differently. Since the MAM is considered, the scaling effect is corrected for.

Figure 8 represents the $\sqrt{F}$-plot for the Clip effect. We may observe that the Clip effect is highly significant for all attributes, so assessors are able to discriminate between the products, therefore we do not exclude any attributes from further analysis.

The product pairwise comparisons may be extracted, although, since there are 90 levels in the Clip factor, there are $C_{90}^2 = 4005$ pairwise comparisons, which are indeed hard to interpret. Hence considering a multi-way product structure might simplify the analysis of product differences by comparing product features plus some additional insight into the data might be gained.

5.2.2. multi-way product structure

In the Car data the following three features form the products: Car, Track and SPL. In order to consider a multi-way product structure three main effects Car, Track and SPL and all possible interactions between them need to be considered instead of one Clip effect. From the one-way product analysis we have deduced that the Repetition effect and the Repetition:Clip effect are non-significant, hence option **2-WAY** may be chosen for the **error structure**. All that results in selecting the following modelling controls in the SensMixed:

error structure = 2-WAY

product structure = 3

scaling correction = Yes

Figure 9 represents the sequential chi-squared values (i.e. from the stepwise selection process) for the tests of random-effects for a multi-way product structure case.

It can be seen that the 2-way interactions Car:Participant, Track:Participant and the 3-way interactions Track:Car:Participant are significant for four out of eight attributes, so the participants disagree in scoring the products according to these features for this attributes. The 3-way interactions Track:SPL:Participant and Car:SPL:Participant are significant for six attributes. The 2-way interactions Rep:Participant and SPL:Participant are significant for almost all attributes. Participant is significant for three attributes and Rep is non significant for all attributes For each attribute a reduced random structure is found which includes only the significant random effects according to the likelihood ratio test with the default Type 1 error rate equal to 0.1. The reduced random structure is then used in tests for the fixed effects.

Figure 10 represents the delta-tilde plot. The three-way interaction is significant for all attributes. The Car effect and its interactions with the other effects are highly significant for all attributes, so the Car effect has a high impact on the ability of assessors to discriminate between the products. The heights of the bars corresponding to the Track and SPL effects and the interaction between them are lower than those pertaining to the Car effect, which means that the size of the Track and SPL effects are lower than the size of the Car effect. However for the attribute att1 the effect of SPL is highly significant and the height of the bar is much higher than for the other

attributes, so there is a high impact of the SPL feature on the ability to discriminate between the products for this attribute.

To complement the results of the mixed model analysis of variance, we may look at the least square means and difference of least squares means given by the post-hoc analysis. From Figures 11 we can see which levels of Car are different for the attribute 1. It can be seen that for level 3 of Car we have the highest score for attribute 1. There is no significant difference between levels 1-2 and 4-6 .

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

Table 1: $\sqrt{\chi^2}$-statistics for LRT for random-effects with significance levels for the BO data

	Track:Participant	Car:Participant	SPL:Participant	Track:Car:Participant	Track:SPL:Participant	Car:SPL:Participan
att1	0.68	0.47	22.03***	0.00	76.91***	269.59***
att2	18.90***	20.74***	23.19***	5.92*	0.00	1.57
att3	0.08	24.67***	38.59***	0.00	26.66***	9.79**
att4	19.08***	15.64***	8.89**	0.64	1.76	32.48***
att5	4.34*	40.10***	6.18*	0.31	16.97***	25.35***
att6	1.69	3.88*	3.73	20.41***	30.81***	8.76**
att7	0.00	14.06***	11.70***	10.63**	69.85***	25.46***
att8	2.53	6.29*	5.79*	21.82***	7.60**	23.42***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

6. Discussion and Conclusion

In this paper we have presented the R package SensMixed as an open source implementation for advanced statistical methods within a mixed effects framework. The aim of this new package is to provide an intuitive and easy-to-use graphical interface that enable the sensory practitioner and non-statistician to analyze properly sensory and consumer data, getting important and interpretable information given by different types of plots and tables. The SensMixed package has implemented an automated analysis of mixed-effect models, using the same technique of the lmerTest package, but applied to all attributes simultaneously. The examples have shown that the new SensMixed package can improve the analysis of sensory data, allowing multi-way product structures, incomplete data and complex error structure. The SensMixed package also provides results of the analysis of random and fixed effects presented in tables and plots, including the new delta-tilde plot and the post-hoc product pairwise comparisons. Beyond that, the SensMixed provides a tool to analyze the extended versions of the Mixed Assessor Model, where a possible multi-way product structure can be accounted for together with the 3-way error structure, a novel contribution. All that makes the SensMixed package, together with its application, a very valuable tool for sensory practitioners as it requires no skills in R-programming and provides advanced statistical methods for analyzing sensory and consumer data.

7. Appendix

7.1. Approach for calculating product differences in MAM

Let us specify a linear mixed effects model in the following way:

$$y = X\beta + Zu + \varepsilon \quad u \sim N_q(0, G) \quad \varepsilon \sim N_n(0, R) \tag{7}$$

The mixed assessor model, where the scaling term is added to the model (7), can then be specified in the following way:

$$y = X_{MAM}\beta_{MAM} + Zu + \varepsilon \quad u \sim N_q(0, G) \quad \varepsilon \sim N_n(0, R) \tag{8}$$

where the size of β_{MAM} is the size of β plus number of the coefficients for the individual slopes from the MAM and X_{MAM} is the corresponding design matrix. Note that the random structure is the same in models (8) and (7). Due to the fact that there is a 100% confounding in β_{MAM}, the contrast vector l for testing the product differences cannot be easily obtained and most of the software would not produce the tests. We have taken an approach where still the tests for the product differences can be obtained for MAM. The tests for the product differences are generally obtained using the t-tests, where the t-statistics is calculated in the following way:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l^\top}} \tag{9}$$

where l is a contrast vector and $\hat{C}$ is an estimated variance-covariance matrix of $\hat{\beta}$. Both $\hat{\beta}$ and $\hat{C}$ are estimated from model (7). Similarly, the test for the product differences for model (8) is:

$$t = \frac{l_{MAM}\hat{\beta}_{MAM}}{\sqrt{l_{MAM}\hat{C}_{MAM}l_{MAM}^\top}} \tag{10}$$

where now $\hat{\beta}_{MAM}$ and $\hat{C}_{MAM}$ are estimated from model (8). The numerators in Equations (9 and 11) correspond to the estimates of product differences and therefore should be the same. Whereas l can be easily obtained, it is hard to get l_{MAM} due to non-estimability. But then the following t-statistics can be considered:

$$t = \frac{l\hat{\beta}}{\sqrt{l\hat{C}_{MAM}l^\top}} \tag{11}$$

where $\hat{C}_{MAM}$ contains rows and columns corresponding to the first β coefficients. The confidence intervals are then computed using the following formula:

$$CI = l\hat{\beta} \pm t_{\frac{\alpha}{2}}(\nu) * \sqrt{l\hat{C}_{MAM}l^{\top}}$$

where ν is calculated using the Satterthwaite's method of approximation based on MAM in Equation 8

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. URL: `http://CRAN.R-project.org/package=lme4` r package version 1.0-4.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. URL: `http://CRAN.R-project.org/package=lme4` r package version 1.1-7.

Beck, T. K., Jensen, S., Bjoern, G. K., & Kidmose, U. (2014). The masking effect of sucrose on perception of bitter compounds in brassica vegetables. *JOURNAL OF SENSORY STUDIES*, *29*, 190–200. doi:`10.1111/joss.12094`.

Brockhoff, P. B., Schlich, P., & Skovgaard, I. (2015). Taking individual scaling differences into account by analyzing profile data with the mixed assessor model. *Food Quality and Preference*, *39*, 156–166.

Brockhoff P. B., L. R. R. K. A. C. R. H. B., Amorimb I. (2015). d-prime interpretation of standard linear mixed model results. *Food Quality and Preference*, .

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2015). *shiny: Web Application Framework for R*. URL: `http://CRAN.R-project.org/package=shiny` r package version 0.11.1.

Jaeger, S. R., Mielby, L. H., Heymann, H., Jia, Y., & Frst, M. B. (2013). Analysing conjoint data with ols and pls regression: a case study with wine. *Journal of the Science of Food and Agriculture*, *93*, 3682–3690. URL: `http://dx.doi.org/10.1002/jsfa.6194`. doi:`10.1002/jsfa.6194`.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2013a). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).*. URL: `http://CRAN.R-project.org/package=lmerTest` r package version 2.0-22.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2013b). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. URL: `http://CRAN.R-project.org/package=lmerTest` r package version 2.0-0.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2013c). *SensMixed: Mixed effects modelling for sensory and consumer data*. R package version 2.0-6.

Kuznetsova, A., Christensen, R. H., Bavay, C., & Brockhoff, P. B. (2015). Automated mixed {ANOVA} modeling of sensory and consumer data. *Food Quality and Preference*, *40, Part A*, 31 – 38. URL: http://www.sciencedirect.com/science/article/pii/S0950329314001724. doi:http://dx.doi.org/10.1016/j.foodqual.2014.08.004.

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food*. Springer Science+Business Media, LLC.

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference*, *2*, 187 – 199. URL: http://www.sciencedirect.com/science/article/pii/095032939090023N. doi:http://dx.doi.org/10.1016/0950-3293(90)90023-N.

Nofima Mat, N., Ås (2008). Panelcheck software. URL: www.panelcheck.com.

Tomic, O., Forde, C., Delahunty, C., & Ns, T. (2013). Performance indices in descriptive sensory analysis a complimentary screening tool for assessor and panel performance. *Food Quality and Preference*, *28*, 122 – 133. URL: http://www.sciencedirect.com/science/article/pii/S0950329312001267. doi:http://dx.doi.org/10.1016/j.foodqual.2012.06.012.

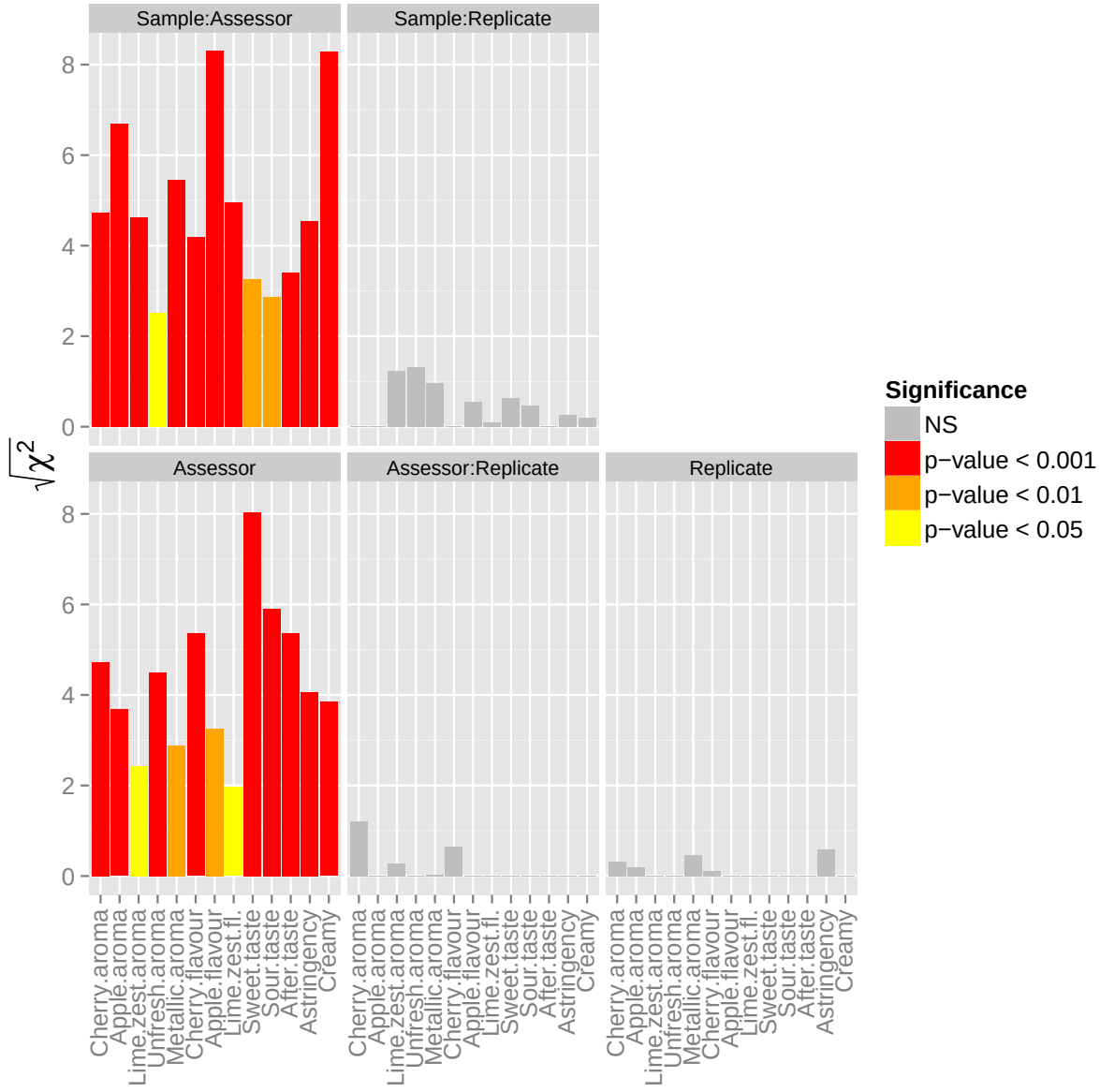


Figure 1: Barplots for $\sqrt{\chi^2}$ of likelihood ratio test for random-effects for the Cherry data

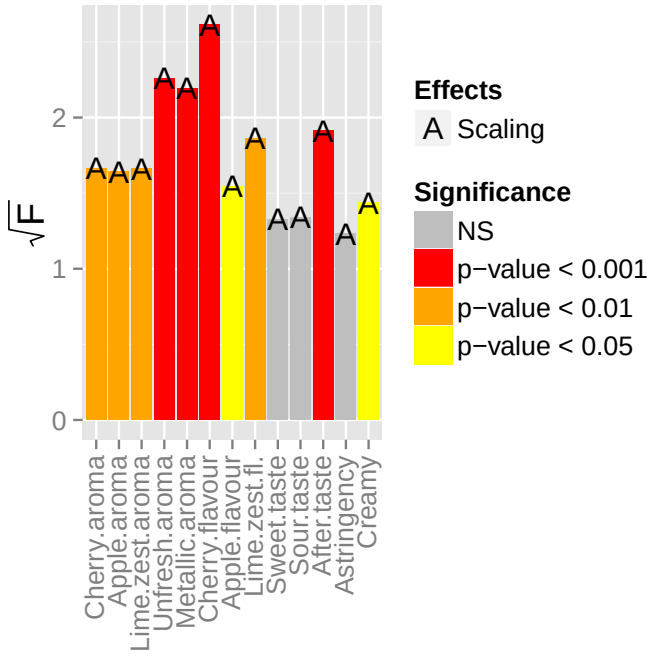


Figure 2: Barplots for $\sqrt{F}$-statistics for the scaling effect for the Cherry data

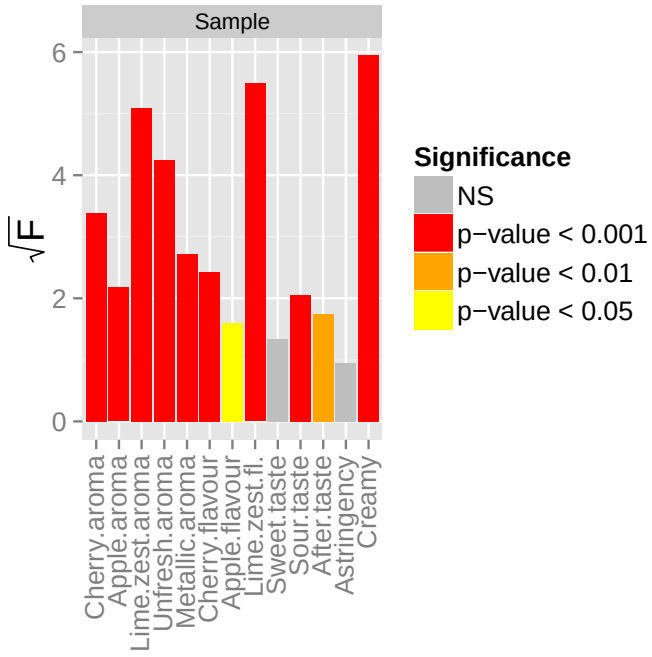


Figure 3: Barplots for $\sqrt{F}$-statistics for fixed-effect Sample for the Cherry data.

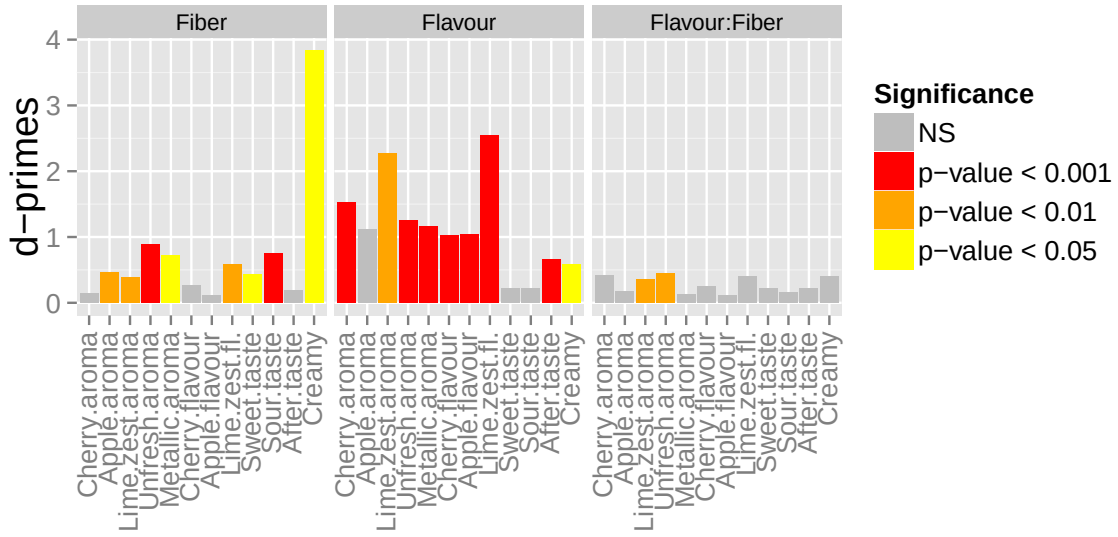


Figure 4: Barplots for delta-tilde estimates for fixed-effects for the Cherry data.

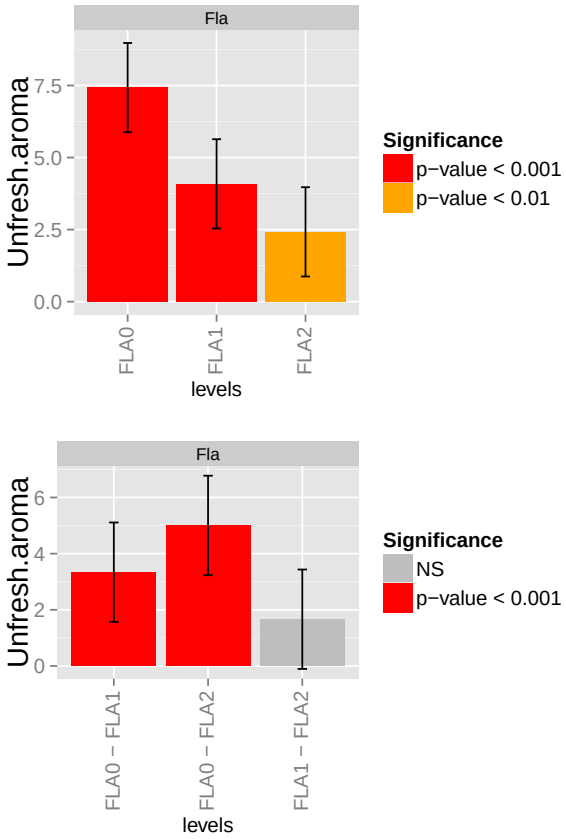


Figure 5: Barplots for least squares means and differences of least squares means for Flavour effect together with 95% confidense intervals for the unfresh aroma attribute

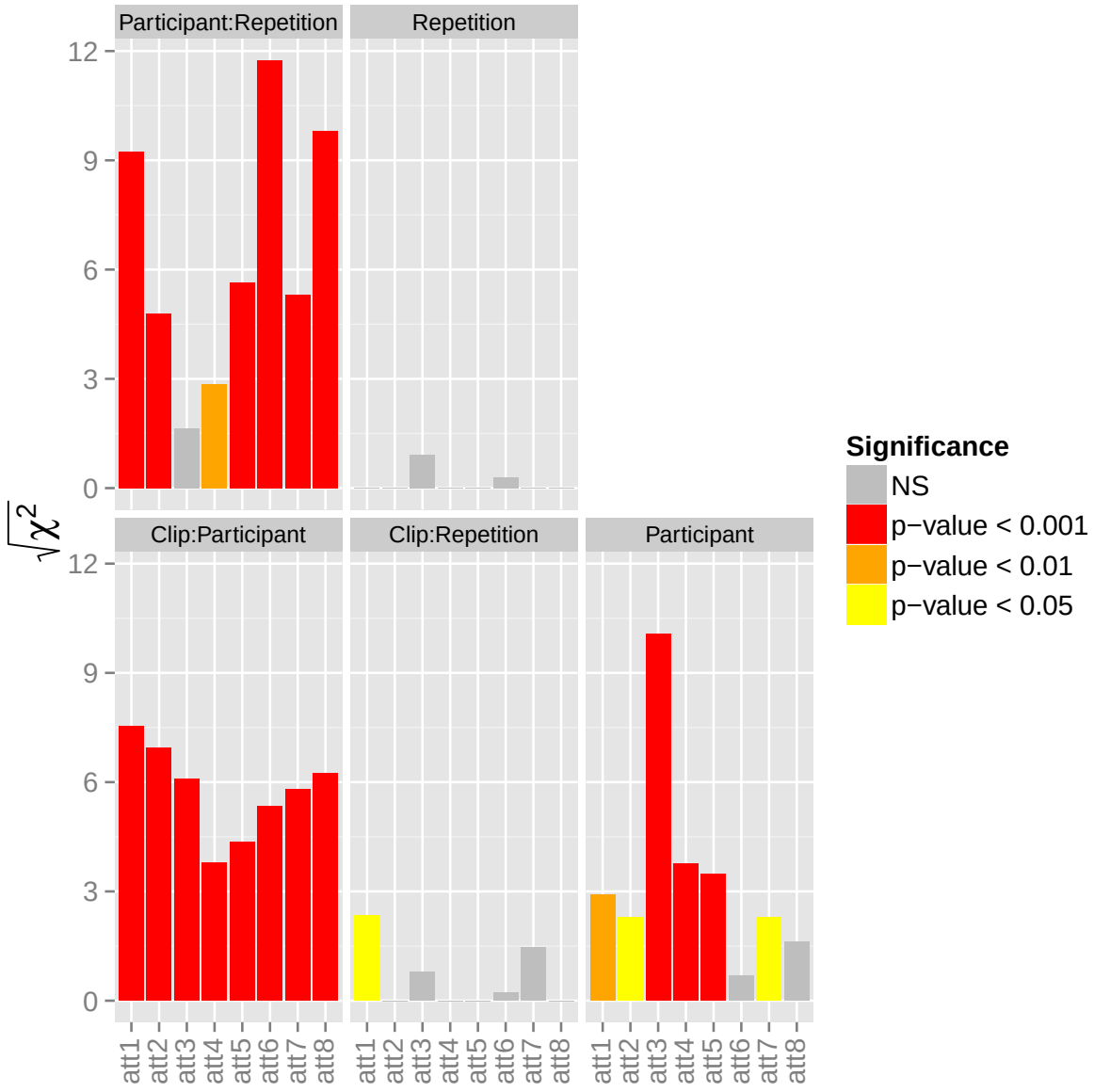


Figure 6: Barplots for the $\sqrt{\chi^2}$ of likelihood ratio test for random-effects for the car audio system data

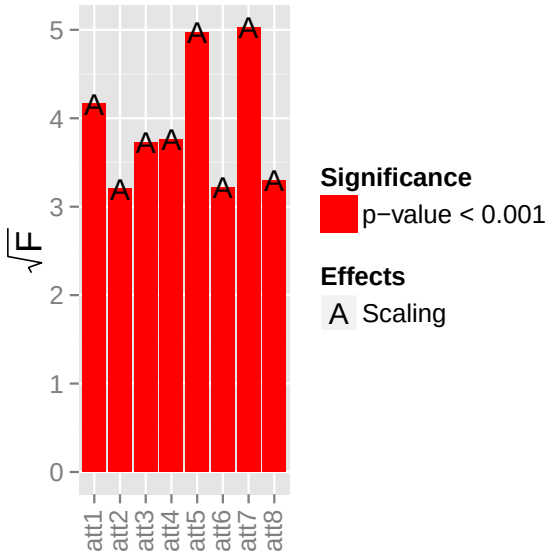


Figure 7: Barplots for $\sqrt{F}$-statistics for fixed-effect Scaling for the car audio system data

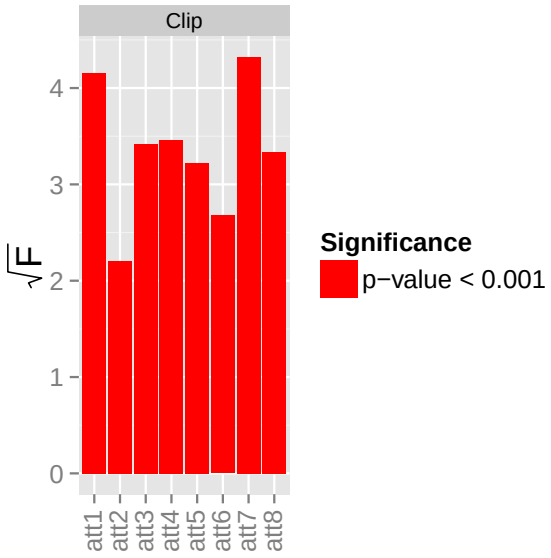


Figure 8: Barplots for $\sqrt{F}$-statistics for fixed-effect Clip for the car audio system data. Corrected for the scaling effect.

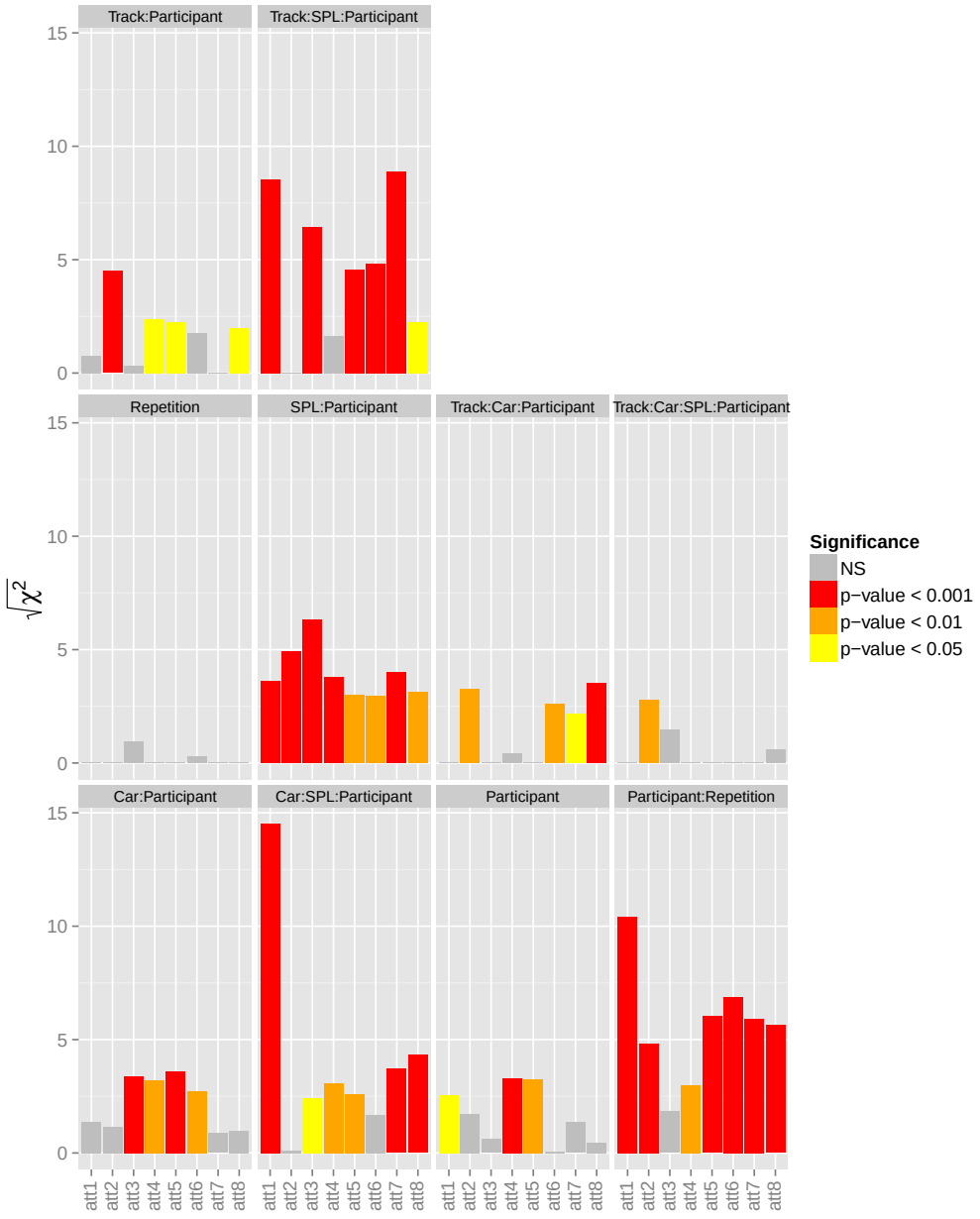


Figure 9: Barplots for $\sqrt{\chi^2}$-statistics of likelihood ratio test for random-effects for the Car Audio System data.

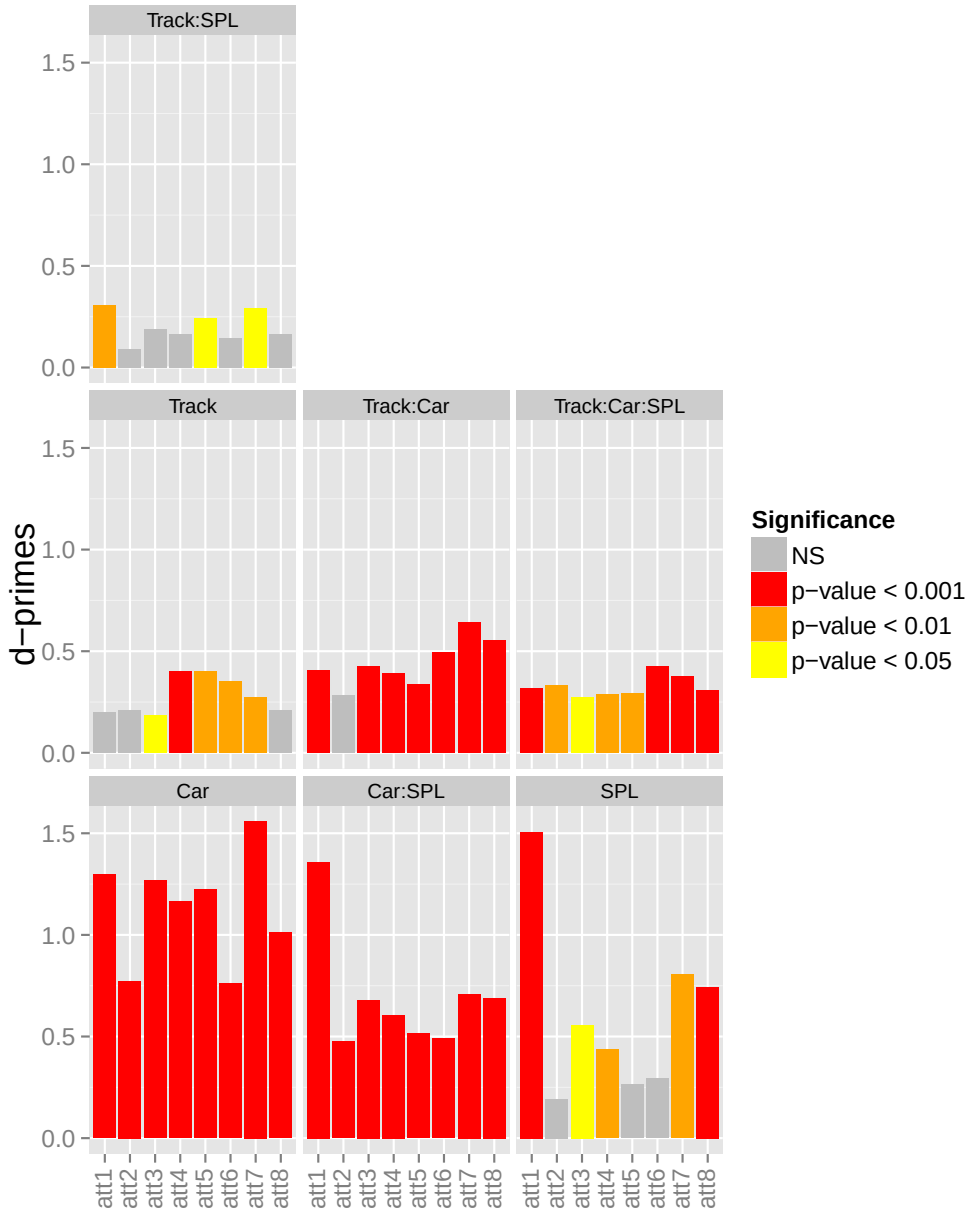


Figure 10: Barplots for delta-tilde estimates for fixed-effects for the Car Audio System data.

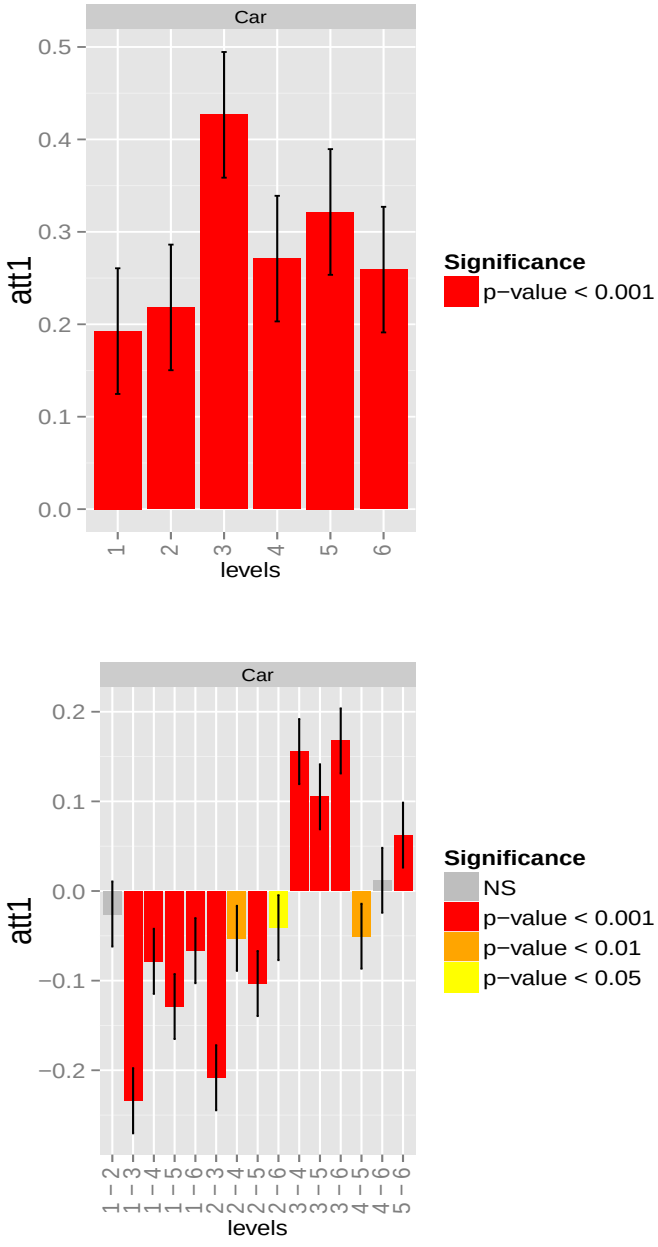


Figure 11: Barplots for least squares means and differences of least squares means for Car effect together with 95% confidense intervals for the attribute 1

Table 2: Overview of the nine cherry fruit drinks used in the study. Sample corresponds to a product factor with 9 levels. Flavour and Fiber are factors corresponding to features of the products each having 3 levels.

Sample	Flavour (addition of lime flavour)	Fibre (addition of β-glycan)
Fla0Fib0	Fla0 (None)	Fib0 (None)
Fla0Fib1	Fla0 (None)	Fib1 (Low)
Fla0Fib2	Fla0 (None)	Fib2 (High)
Fla1Fib0	Fla1 (Low)	Fib0 (None)
Fla1Fib1	Fla1 (Low)	Fib1 (Low)
Fla1Fib2	Fla1 (Low)	Fib2 (High)
Fla2Fib0	Fla2 (High)	Fib0 (None)
Fla2Fib1	Fla2 (High)	Fib1 (Low)
Fla2Fib2	Fla2 (High)	Fib2 (High)

Table 3: Likelihood ratio tests for the random-effects and their order of elimination representing Step 1 of the automated analysis for the attribute Sweet.taste

	Chi.sq	Chi.DF	elim.num	p-value
Sample:Assessor	10.64	1	0	0.001
Sample:Replicate	0.39	1	1	0.530
Assessor	64.60	1	0	<0.001

Table 4: F-tests for the fixed-effects for the attribute Sweet.taste

	Sum Sq	Mean Sq	NumDF	DenDF	F-value	d-tilde	Pr(>F)
Sample	107.94	13.49	8	63.00	1.80	0.48	0.094
Scaling	119.07	13.23	9	63.00	1.76	1.00	0.094

Table 5: F-tests for the fixed-effects for the attribute Astringency

	Sum Sq	Mean Sq	NumDF	DenDF	F-value	d-tilde	Pr(>F)
Sample	79.15	9.89	8	63.00	0.91	0.38	0.516
Scaling	149.09	16.57	9	63.00	1.52	1.00	0.160

Table 6: $\sqrt{\chi^2}$-statistics for LRT for random-effects with significance levels for the Cherry data

	Flavour:Assessor	Fiber:Assessor	Flavour:Fiber:Assessor	Assessor
Cherry.aroma	1.64	0.24	22.44^{***}	22.34^{***}
Apple.aroma	23.70^{***}	0.81	4.88^{*}	0.00
Lime.zest.aroma	48.57^{***}	0.08	1.17	0.00
Unfresh.aroma	0.00	0.00	6.27^{*}	20.29^{***}
Metallic.aroma	0.71	8.16^{**}	7.95^{**}	0.87
Cherry.flavour	0.00	4.75^{*}	5.31^{*}	8.63^{**}
Apple.flavour	0.00	13.21^{***}	20.78^{***}	1.28
Lime.zest.fl.	0.00	0.09	24.64^{***}	3.86^{*}
Sweet.taste	0.05	0.00	10.64^{**}	64.60^{***}
Sour.taste	0.10	0.00	8.19^{**}	34.95^{***}
After.taste	0.00	0.22	11.58^{***}	28.83^{***}
Creamy	0.00	12.29^{***}	18.45^{***}	0.01

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Appendix D

ConsumerCheck: a software for analysis of sensory and consumer data

Tomic, O., Brockhoff, P. B. , **Kuznetsova, A.**, Naes, T.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

ConsumerCheck: a software for analysis of sensory and consumer data

Oliver Tomic
Norwegian Knowledge Centre for the Health Services

Alexandra Kuznetsova
Technical University of Denmark

Per Bruun Brockhoff
Technical University of Denmark

Thomas Graff
TGXnet

Tormod Næs
Nofima 1
University of Copenhagen 2

Abstract

ConsumerCheck is a software for statistical analysis of data from sensory and consumer science. ConsumerCheck provides an intuitive and easy-to-use graphical user interface to a number statistical methods that are often used in the field of sensometrics. The data that are to be analysed are typically acquired from consumer trials and from descriptive analysis / sensory profiling that was performed by trained sensory panels. Besides some simple descriptive statistics the main statistical methods implemented in ConsumerCheck are principal component analysis, preference mapping, partial least squares regression, principal component regression and conjoint analysis, all well established methods in the field of sensometrics and available in several commercial software packages. The ConsumerCheck software is a by-product of an international research project and the aim for developing the software was to provide an open source alternative that makes the implemented statistical methods widely available to the public at no cost.

Keywords: consumer liking data, descriptive analysis data, sensory profiling data, principal component analysis, PCA, preference mapping, partial least square regression, PLS, principal component regression, PCR, Conjoint analysis, Python, R, **SeneMixed**, **lmerTest**.

1. Introduction

In sensory and consumer science (Lawless and Heymann 2010) various types of methods exist for measurement of human responses triggered by sensory stimuli such as taste, smell,

touch, etc. These methods generate data that need to be analysed with appropriate statistical methods (Næs, Brockhoff, and Tomic 2010; Martens and Martens 2001) in order to learn more about the consumers, their sensory preferences and their buying and consumption habits. Many of these statistical methods have been long available to the public, either through proprietary software with polished graphical user interfaces (GUI) or as part of free statistical packages in open source programming languages such as Python or R. The open source option is attractive to many since there are virtually no costs to the user for acquiring the software and using it for analysis of their data. Unfortunately, those open source packages often require some minimum level of programming skills from the user such that he or she would be able to apply the implemented methods for data analysis. There are several useful R packages such as **SensMixed** (Kuznetsova, Bruun Brockhoff, and Haubo Bojesen Christensen 2013b), **sensR** (Christensen and Brockhoff 2014), **SensoMineR** (Husson, Le, and Cadoret 2014) that contain most of the standard sensometrics methods for analysis of sensory and consumer data. Some of them may be even accessed through a general GUI like **Rcmdr**. However, there are a number of users that would still prefer a GUI that makes the use of the statistical methods more intuitive. ConsumerCheck tries to address this issue by providing a GUI that is tailored towards each of the implemented statistical methods and as such makes them easy to apply. Besides some simple descriptive statistics methods such as histograms and box plots the main statistical methods implemented in ConsumerCheck are principal component analysis (PCA), preference mapping based on partial least squares regression (PLSR) and principal component regression (PCR) as well as conjoint analysis. The ConsumerCheck GUI is inspired by PanelCheck (Tomic, Luciano, Nilsen, Hyldig, Lorensen, and Næs 2010; Tomic, Nilsen, Martens, and Naes 2007), a well established open source software (DOI PanelCheck) within the field of sensometrics for performance analysis of trained sensory panels which has been available to the public since 2006.

Although ConsumerCheck origins from the field of food science it may also be applied to data from non-food domains that produce data with the same structure. Some examples would be entertainment electronics industry, car industry or different types of services. ConsumerCheck is therefore broadly applicable to any areas where sensory stimuli are measured. Moreover, with the rather generic statistical methods PCA, PLSR and PCR, ConsumerCheck can be applied for many other types of measurement data that are not based on sensory stimuli, such as chemical or physical measurement data. ConsumerCheck is the result of a international research project (2009 - 2013) that was funded by project participants of the Danish and Norwegian industry, The Research Council of Norway and The Danish AgriFish Agency.

2. Type of data and their properties

ConsumerCheck was initially designed for analysis of four types of data that are common in sensory and consumer science, i.e. *consumer liking* data, *consumer characteristics* data, *product design* data and *descriptive analysis / sensory profiling data* (for more details see section 2.1 through 2.4). However, ConsumerCheck contains statistical methods that are generic, such as PCA, PLSR and PCR that allow for analysis of data data from any domain (see section 2.7), not only from sensory and consumer science. The user only needs to keep in mind that the implemented statistical methods can analyse only data that are suitable for the method. For illustrational purposes five data sets from two sensory experiments, *hams* and *apples* respectively, will be used and analysed in this paper. More information on the *hams*

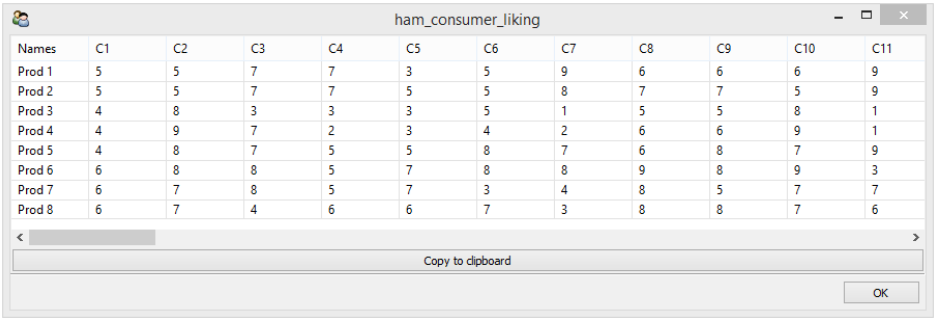


Figure 1: This is a screenshot showing a small part of the *ham consumer liking* data that are described in section 2.6. The data window shows product ratings of the first eleven consumers (C1 to C11) for the eight tested products.

and *apples* data are found in section 2.6. It is important to note that as of ConsumerCheck version 1.2.0 all data (except for column and row names) need to be numerical. If categories are to be used for factors in conjoint analysis (see section 3.6 and 6.7) one should use numbers as factor levels instead of strings or characters.
Moreover, one needs to keep in mind that missing values are allowed in general, but not all implemented statistical methods can handle them. As of version 1.2.0 of ConsumerCheck, only the *Conjoint* method can handle missing data and provides results when starting off a data set with missing values. The other methods cannot handle missing values and will provide an error message when attempting to carry out computations. More information on how missing values are handled in ConsumerCheck are provided in section 6.1.4.

2.1. Consumer liking data

Consumer liking data X_{cl} are acquired through consumer trials where each consumer rates his or her liking of a product on a hedonic scale (typically from 1 to 5, 1 to 7 or 1 to 9, where the 1 represents "do not like at all" and the highest value represents "like very much"). The dimension of a *consumer liking* data is $(J \times N)$ where the $j = 1 \dots J$ objects (products) are represented by rows and $n = 1 \dots N$ variables (consumers) are represented by columns. Fig. 1 shows what *consumer liking* data may look like.

Descriptive statistics and visualisation of the *consumer liking* data distribution may be obtained with the methods implemented in *Basic stat liking* (see section 3.1). Moreover, *consumer liking* data can be analysed by using PCA (see sections 3.2 and 6.4); by using Preference mapping in combination with *descriptive analysis / sensory profiling* data (see sections 3.3 and 6.5); by using PLSR/PCR in combination with either *product design* data or *consumer characteristics* data (see section 3.4, 3.5 and 6.6); by using conjoint analysis (see section 3.6 and 6.7) together with *consumer characteristics* data and *product design* data.

2.2. Consumer characteristics data

Consumer characteristics data X_{cc} are data that provide background information on the consumers that have participated in the consumer trial. The *consumer characteristics* data are of dimension $(N \times I)$ where the $n = 1 \dots N$ objects (consumers) are represented by rows and $i = 1 \dots I$ variables (consumer characteristics variables) are represented by columns. With

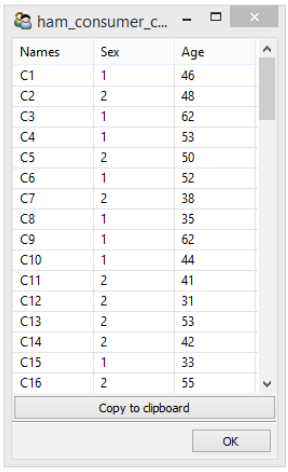


Figure 2: This is a screenshot showing a small part of the *ham consumer characteristics* data that are described in section 2.6. The data window shows two *consumer characteristics* variables, i.e. sex and age of the first 16 consumers (C1 to C16).

the current version (1.2.0) of ConsumerCheck the consumer characteristics variables can be of any type, such as gender, age, country of origin, income, size of household, habits, etc as long as their levels are represented by integers. Note that the more levels a characteristics variable consists of the longer computation times will be. Although there is no limit to how many levels are allowed in a categorical variable we recommend to limit them to about five to six to keep computation times within reasonable limits. *Consumer characteristics* data are usually analysed with conjoint analysis (see section 3.6 and 6.7) together with *consumer liking* data and *product design* data. There is no limit to how many characteristics variables the *consumer characteristics* data may consist of, but one should not include more than three to four in the conjoint model, because computation time and complexity of the model would increase much. Instead, several models with fewer variables should be run to identify important characteristics. Fig. 2 shows what consumer characteristics data may look like.

Consumer characteristics data can be analysed with PCA (see sections 3.2 and 6.4) provided that there are at least three variables. It is important to note that when analysing *consumer characteristics* data all variables should be standardised since background variables are typically of different nature using different scales or units. Furthermore, it should be noted that it is not meaningful to include categorical variables (such as sex, where for example male is coded as 1 and female is coded as 2) and mix them with continuous variables (such as age) when applying PCA to the consumer characteristics data. Extended versions of PCA that handle this type of situation are available, such as the R package **PCAmixdata** (Chavent, Kuentz-Simonet, Labenne, and Saracco 2014), but this is not supported by ConsumerCheck at its current version.

2.3. Product design data

If the products rated by the consumers were produced by use of an experimental design, then these design data can be imported into ConsumerCheck and utilised for statistical analysis. The product design data X_d are of dimension $(J \times M)$ where $j = 1 \ldots J$ objects (products)

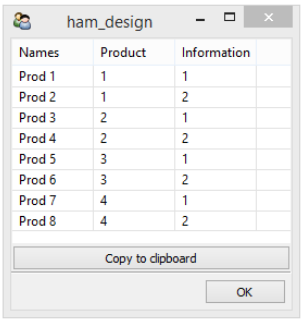


Figure 3: This is a screenshot showing the product design of the *ham* data that are described in section 2.6. The data window shows design variables *Product* and *Information* for the eight products. For more details on the *ham* data see section 2.6

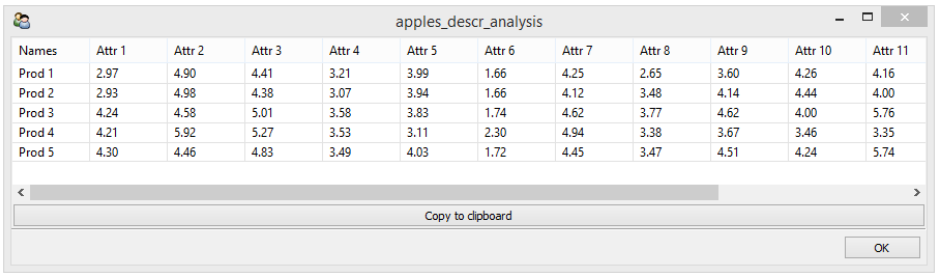


Figure 4: This is a screenshot showing a part of the *apples descriptive analysis / sensory profiling* data that are described in section 2.6. The data window shows intensity ratings for the first eleven sensory attributes for the five apple products that were tested by the trained sensory panel.

are represented by rows and $m = 1 \dots M$ design variables are represented by columns.

The product design may be of type *full factorial* or *fractional factorial*. The product design data may be used for analysis in PLSR or PCR (section 3.4, 3.5 and 6.6) or in conjoint analysis (see section 3.6 and 6.7). Fig. 3 shows what product design data may look like.

2.4. Descriptive analysis or sensory profiling data

Descriptive analysis (often also referred to as *sensory profiling*) is a standard sensory tool that has an important role in research and product development (Lawless and Heymann 2010). When performing descriptive analysis a panel of trained assessors rate for each tested product the perceived intensity of defined sensory attributes on scales. The *descriptive analysis / sensory profiling* data X_{da} are of dimension $(J \times K)$ where the $j = 1 \dots J$ objects (food products) are represented by rows and $k = 1 \dots K$ variables (sensory attributes) are represented by columns. Fig. 4 shows what *descriptive analysis / sensory profiling* data may look like.

Descriptive analysis / sensory profiling data can be analysed with PCA (see sections 3.2 and 6.4); with preference mapping (see sections 3.3 and 6.5) together with *consumer liking* data; with PLSR or PCR (see section 3.4, 3.5 and 6.6) together with either *product design* data or *consumer characteristics* data.

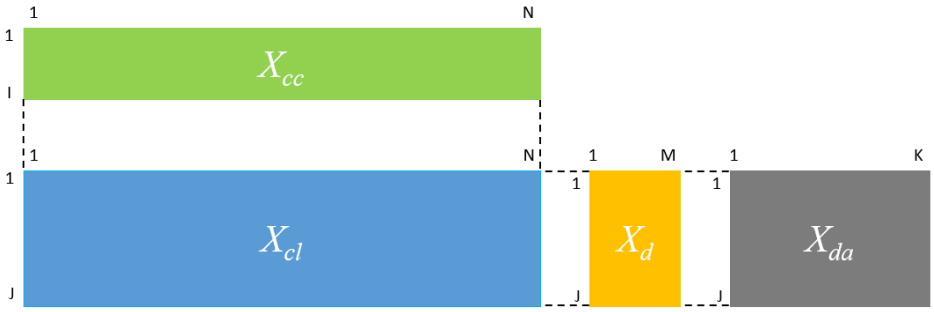


Figure 5: This plot illustrates how the four data types described above relate to each another. *Consumer liking data* X_{cl}, *product design matrix* X_d and *descriptive analysis / sensory profiling data* X_{da} share the common axis or dimension of J products. *Consumer liking data* X_{cl} and *consumer characteristics data* X_{cc} share the common axis or dimension of N products.

2.5. Relationship between the four types of sensory and consumer data

Fig. 5 shows how the *consumer liking* data X_{cl}, *consumer characteristics* data X_{cc}, *product design* matrix X_d and *descriptive analysis / sensory profiling* data X_{da} relate to each other. Note that for illustration purposes *Consumer liking* X_{cl} is plotted transposed compared to how it is organised prior to import into ConsumerCheck. For X_{cl}, X_d and X_{da} the common axis are the tested products. For X_{cl} and X_{cc} the common axis are the consumers.

2.6. Real world data used in examples

In order to illustrate how to apply the statistical methods implemented in ConsumerCheck a number of data sets are used that were acquired through two independent sensory and consumer science experiments.

Apple data

The apple data consist of two data matrices: (I) a data matrix of type *consumer liking* where 108 consumers have rated 5 apples. Hence its dimension is (5×108); (II) a data matrix of type *descriptive analysis / sensory profiling* where the same 5 apples were described by a trained sensory panel using 14 attributes. Hence its dimension is (5×14).

Ham data

The ham data consist of three data matrices: (I) a data matrix of type *consumer liking* (see Fig. 1) where 81 consumers rated the four hams twice (presented once as Norwegian and once as Spanish ham; see more details below in (III)). Hence its dimension of (8×81); (II) a data matrix of type *consumer characteristics* (see Fig. 2) consisting of two variables (named *Sex* and *Age*) that provide background information on the 81 consumers. Hence its dimension of (81×2); (III) a matrix of type *product design* consisting of two design variables (see Fig. 3). The first design variable is named *Product* and represents the four hams that were presented the consumers to rate their liking. This means that there are four levels for design variable *Product*. As part of the experiment each of the four hams were presented to the consumers twice, once pretending they were Norwegian ham and once pretending they were Spanish

ham. The aim was to find out whether the country of produce would influence the liking of the consumer. This is determined by the second design variable, named *Information*. It has two levels, where *1* indicates that the ham was presented as Norwegian ham and *2* indicates that it was presented as Spanish ham. From the two design variables we get a (4×2) full factorial experimental design which results in a total of eight "unique" ham products named *Prod 1* through *Prod 8*. The *product design* data therefore is of dimension (8×2) where each row represent a unique combination of levels from the two design variables *Product* and *Information*.

2.7. Other data types

As mentioned above, ConsumerCheck was initially designed for analysis of data from sensory and consumer science (see section 2.1 through 2.4). However, since some statistical methods are generic there is no reason to limit the use of ConsumerCheck to only data from sensory and consumer science. Any kind of data that are suitable for analysis with PCA (section 3.2), PLSR (section 3.4) and PCR (section 3.5) may be imported to ConsumerCheck and be tagged as type *Other* when importing them.

3. Statistical methods in ConsumerCheck

ConsumerCheck contains a number of statistical methods that are very common in analysis of sensory and consumer data. As of version 1.2.0 the following methods are implemented:

- standard statistical methods for obtaining descriptive statistics from *consumer liking* data such as box plots and histograms, see section 3.1
- principal component analysis (PCA), see section 3.2
- preference mapping (prefmap), see section 3.3
- partial least squares regression (PLSR), see section 3.4
- principal component regression (PCR), see section 3.5
- conjoint analysis, see section 3.6

All methods are thoroughly described in textbooks and scientific papers, which is why we keep the methods sections short and discuss in detail only issues that directly relate to the use of ConsumerCheck. Further on in this paper, the graphical user interface (GUI) for each method is discussed in detail (section 6.3 through 6.7), including how to set model parameters, how to obtain results and how plots and tables are interpreted.

3.1. Basic statistics for consumer liking data

Under the *Basic stat liking* tab (see Fig. 10) there are three types of plots available for quick inference the *consumer liking* data. The first two, that is the *Box plot* and *Stacked histogram*, visualise the distributions of the liking ratings across all consumers for each of the tested products or across all tested products for each consumer. The third type, the

Single product histogram, visualises the distribution of the ratings across all consumers for one specific product at the time in an ordinary histogram.

Box plots

The *box plot* (for an example see Fig. 11) describes how the consumer liking rates are distributed for each product. More precisely, it shows how the ratings are distributed between the 25 and 75 percentile of the data. The dark green line across the box indicates the median value (CHECK THIS). The vertical lines above and below a box indicate which range of the scale was used. More practical details on how to generate box plots in ConsumerCheck and the interpretation of results are found in section 6.3.

Stacked histograms

Stacked histograms are another way of visualising the consumer liking rates for each product (see Fig. 12). Here, however, one can see for every product how often each liking rate was used. More practical details on how to generate stacked histograms in ConsumerCheck and interpretation of results are found in section 6.3.

Single product histogram

These are ordinary histograms showing the distribution of ratings across all consumers for a single product. More practical details on how to generate histograms in ConsumerCheck and how to interpret results are provided in section 6.3.

3.2. Principal component analysis

PCA (Mardia, Kent, and Bibby 1979) as implemented in ConsumerCheck is coded in Python and uses the NIPALS algorithm (Wold 1982) to provide scores, loadings, correlation loadings, calibrated and validated explained variances for the analysed data. Furthermore, one can access predicted (i.e. reconstructed) versions of the analysed data after each PC for both calibration and validation. PCA is accessible through the *PCA* tab (see Fig. 14 and details on the usage is provided in section 6.4). If needed, further computation results, such as root mean square error of calibration and cross validation (RMSEP and RMSECV, etc.), are available from the PCA class when using the Python source code directly outside ConsumerCheck. The PCA implemented in ConsumerCheck contains an option for variable standardisation if equal weight is to be given to each variable. It is important to note that ConsumerCheck automatically leaves out variables with zero variance when standardisation of variables is selected since the standard deviation for such a variable is STD=0. Whenever this happens, ConsumerCheck provides information on which variables have been left out in a message box dialog. The calibrated explained variance provided by the PCA model describes how much of the total variance in the data is explained by each principal component (PC). The cumulative calibrated explained variance (see Fig. 19 for an example) increases with every PC added to the model. The validated explained variance is computed by systematically leaving out objects/rows from the data, then computing new PCA models and using the new loadings to predict values of the data that were left out. The closer the predictions of the left out data are to the real values of the left out data, the more robust the model. Note that the validated explained variance is computed using full cross validation, also known as leave-one-

out in other scientific fields. Currently, for user friendliness and simplicity reasons there are no options to change this setting, but future versions of ConsumerCheck may provide k-fold cross validation. But for most of the practical cases in sensory and consumer analysis full cross validation should be sufficient, since the number of objects or products measured is usually low and the products typically independent. More detailed information on calibrated and validated explained variances are found elsewhere (Martens and Næs 1989).

3.3. Preference mapping

Preference mapping (Greenhoff and MacFie 1994; McEwan 1996) is a much used statistical method in the field of sensometrics that analyses *consumer liking* and *descriptive analysis / sensory profiling* data together. It is available through the *Prefmap* tab. Preference mapping visualises individual differences between consumers and their preference for products with certain sensory attributes. The preference mapping model is actually a regression model that consists of an X and Y matrix and that attempts to find components that describe common variation between the two. Depending on whether the *consumer liking* data or *descriptive analysis / sensory profiling* data is chosen to be the X matrix, one speaks of internal or external preference mapping (Næs *et al.* 2010), respectively. Furthermore, for the computation of the components one can choose between partial least square regression (PLSR) and principal component regression (PCR). Both PLSR (see section 3.4) and PCR (see section 3.5) are well established multivariate regression methods in the field of sensometrics. Having the option to choose between the two can be seen as if these were two different "engines" that power the computations of the preference mapping model. The impact of making a choice between internal or external preference mapping combined with the selection between either PLSR and PCR is discussed elsewhere (Næs *et al.* 2010). Preference mapping and its "engines" PLSR and PCR are coded in Python. Details on the usage of preference mapping through its GUI is provided in section 6.5. One can access X scores, X loadings, Y loadings, calibrated and validated explained variances for X and Y. Using the Python source code one can access also results such as root mean square error of calibration and cross validation (RMSEP and RMSECV, etc.). As with PCA (see section 3.2), the calibrated explained variance is computed from the full set of objects/rows in X an Y, whereas the validated explained variance is computed by use of full cross validation.

3.4. Partial least squares regression

PLSR (Wold 1982) is a multivariate regression method that is frequently used in the field of sensometrics. The main purpose of the method is to find components that describe common variation between two data matrices X and Y. In ConsumerCheck the NIPALS algorithm (Wold 1982) is applied to compute results for PLSR. It searches for components by iterating forth and back between X and Y, which means that both X and Y simultaneously influence the compuation of components unlike with PCR (see section 3.5) where only X determines the components.

PLSR as implemented in ConsumerCheck is coded in Python and provides X scores, X loadings, Y loadings, X & Y correlation loadings as well as calibrated and validated explained variances for X and Y. If needed, further computation results, such as root mean square error of calibration (RMSEP) and cross validation (RMSECV), etc., are available from the PLSR class when using the Python source code directly outside ConsumerCheck.

PLSR is accessible through the *Prefmap* tab (see Fig. 20 and details on the usage in section 6.5) and the *PLSR/PCR* tab (see Fig. 29 and details on the usage in section 6.6). Under the *Prefmap* tab the use of PLSR is restricted to only *consumer liking* data and *descriptive analysis / sensory profiling* data, since *preference mapping* deals only with these to types of data. Under the *PLSR/PCR* tab other types of data may be analysed as for example *consumer characteristics* data together with transposed *consumer liking* data or *product design* data together with *consumer liking* data. If available, other types of data may be analysed with PLSR, either together with any of the four data types described from section 2.1 through 2.4 or separately.

3.5. Principal component regression

PCR (Martens and Næs 1988) is another multivariate regression method that is well established in the field of sensometrics. PCR is basically a two-step procedure. First, PCA is applied to the X matrix, finding principal components that explain the variance in the X data only. Second, linear regression is applied to project the variables of Y onto the PCA subspace of X. In this way the resulting components are influenced by the variation in X only, unlike PLSR (see section 3.4) where both X and Y influence the computation of components. PCR as implemented in ConsumerCheck is coded in Python and provides X scores, X loadings, Y loadings, X & Y correlation loadings as well as calibrated and validated explained variances for for X and Y. If needed, further computation results, such as root mean square error of calibration (RMSEP) and cross validation (RMSECV), etc., are available from the PLSR class when using the Python source code directly outside ConsumerCheck.
As with PLSR above, PCR is accessible through the *Prefmap* tab (see Fig. 20 and details on the usage in section 6.5) and the *PLSR/PCR* tab (see Fig. 29 and details on the usage in section 6.6). Under the *Prefmap* tab the use of PCR is restricted to only *consumer liking* data and *descriptive analysis / sensory profiling* data, since *preference mapping* deals only with these to types of data. Under the *PLSR/PCR* tab other types of data may be analysed as for example *consumer characteristics* data together with transposed *consumer liking* data or *product design* data together with *consumer liking* data. If available, other types of data may be analysed with PLSR, either together with any of the four data types described from section 2.1 through 2.4 or separately..

3.6. Conjoint analysis

Conjoint analysis (Green and Rao 1971; Green and Srinivasan 1978) is a method for analysing the effects of *design factors* (which are stored in the *product design* matrix; see section 2.3) and *consumer characteristics* (see section 2.2) on *consumer likings* (see section 2.1). A common approach is to analyse it in a mixed effects model framework, where random effects consist of consumer effect and interactions between consumer effects and design factors, and fixed effects consist of design factors and consumer characteristics and possibly interactions between them.
Hence, in this type of analysis the following data set types are used: *product design* matrices, *consumer liking* matrices as well as *consumer characteristics* matrices. Mixed effects models in conjoint analysis in ConsumerCheck are constructed using the R package **lme4** (Bates,

Maechler, Bolker, and Walker 2014). The tests and post-hoc analysis for the models are performed using the **lmerTest** R package (Kuznetsova, Bruun Brockhoff, and Haubo Bojesen Christensen 2013a). Conjoint Analysis as implemented in ConsumerCheck has a number of nice features: it can handle unbalanced data, multiple crossed effects, it can automatically find parsimonious models and perform post-hoc analysis. Different degrees of complexity (structure 1, 2 and 3) can be chosen by the user. More practical details on how to generate conjoint related plots in ConsumerCheck and how to interpret results are given in section 6.7.

3.7. Future implementations of statistical methods

ConsumerCheck is an ongoing project and there are plans to extend the ConsumerCheck software with more statistical methods.

4. Software architecture

The application framework and main part of the ConsumerCheck software is programmed in Python. The **numpy** package (Oliphant 2007) was used for coding of the multivariate statistical methods PCA, PLSR and PCR. The **ETS** package (Enthought Tool Suite) was used for building the GUI and the functionality for handling user input and using statistical methods for analysis. Conjoint analysis in ConsumerCheck is implemented with the R package named **lmerTest** which is accessed by the framework through the Python based **PypeR** package that interfaces Python and R.

5. Software installation

At the time of writing of this paper the following installation options for ConsumerCheck 1.1.0 are available:

5.1. Windows platform

Windows users have basically three choices of how to install and run ConsumerCheck. (I) The easiest option is to install ConsumerCheck using the Windows binaries, i.e. standard Windows installation wizard tool (.msi file). Windows binaries are available through the ConsumerCheck website (http://www.consumercheck.co) or from Sourceforge (sourceforge.xx.net). Python, R and other software packages used by ConsumerCheck do not need to be installed prior to installing ConsumerCheck since they are all provided with the Windows binaries. Also, if Python and R are already installed, they will not be affected by ConsumerCheck, since ConsumerCheck has its own ecosystem with own versions of Python and R. (II) The second, but more complicated option for Windows users to run the source code directly. To be able to do so, however, a number of software packages need to be installed first. Section 5.3 lists which software is needed in order to have the source code run properly. (III) The third option is to run ConsumerCheck through a virtual machine (see section xx).

5.2. Mac and Linux platform

Mac and Linux users have two options to run ConsumerCheck. (I) The first and more cumbersome option is to run the source code directly. To be able to do so, a number of software

packages need to be installed first. Section 5.3 lists which software is needed in order to have the source code run properly. (II) The second option is to run ConsumerCheck through a virtual machine (see section xx).

5.3. Running the source code

The ConsumerCheck source code can be downloaded at ConsumerCheck website (http://www.consumercheck.co) or Sourcforge (sourceforge.xx.net). It consists of mainly of Python code and some R code. The source code should run cross-platform, although at the time of writing it has been tested only on Linux and Windows. ConsumerCheck was successfully run on Windows and Linux with the following installations:

Python environment:

- Python 2.7 (not tested on Python 3.X)
- **bbfreeze** 1.1.3
- **chaco** 4.2.0
- **colormath** 1.0.8
- **enable** 4.2.0
- **numpy** 1.8.2
- **openpyxl** 1.7.0
- **pandas** 0.12.0
- **pyface** 4.2.0
- **pyparsing** 1.5.6
- **PypeR** 1.2.0
- **pytest** 2.6.3
- **traits** 4.2.0
- **traitsui** 4.2.0
- **wxPython** 2.8
- **xlrd** 0.9.2

R environment:

- R $>$ 3.0
- **lmerTest** (may require other R packages)

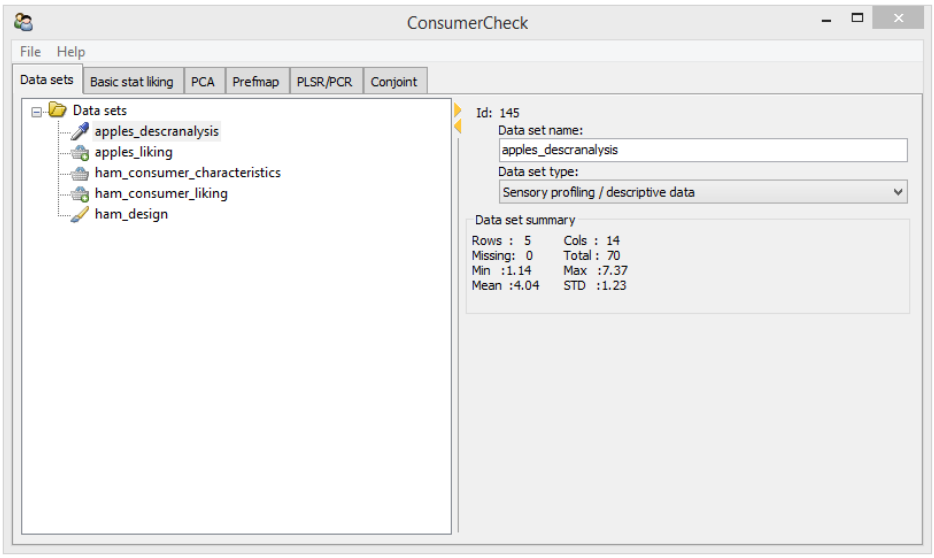


Figure 6: This is a screenshot of the graphical user interface of ConsumerCheck. The screenshot shows five data sets that were previously imported in the last ConsumerCheck session. These are the real world data that were described in section 2.6

At the time of writing Python and most of the Python packages listed above are included in the free Python distribution provided by Continuum Analytics (www.continuum.io). Missing Python packages may be installed from the command line in this way: `pip install PypeR`. To check which Python packages are already installed type the following at the command line: `pip list`. R and R packages are installed in the standard way. (ref R installation)

5.4. Using a virtual machine

6. How to use ConsumerCheck

This section presents the graphical user interface to the reader and discusses all possible settings for data import, data summary and statistical analysis. A brief summary on the statistical methods are provided from section 3.1 through 3.6.

The main widget of the graphical user interface (GUI) is shown in Fig. 6. In order to navigate from one statistical method to another there are a number of tabs at the top of the widget where each tab represents a statistical method except for the first one. The tabs are named: *Data sets, Basic stat liking, PCA, Prefmap* and *Conjoint*. All tabs have the same structure: (I) a so-called tree-control on the left side where the user can generate plots or tables by double clicking on an tree-control item; (II) a panel on the right side where various method-specific parameters can be set. In the following sub sections each tab will be explained in detail. First, however, following the chronological order of data analysis, the data need to be imported.

6.1. Data import

Accepted file formats

ConsumerCheck accepts several file formats for data import.

- plain files such as **.txt** or **.csv**
- Excel files, both **.xls** and **.xlsx**

Note that ConsumerCheck remembers which data sets were imported in the last session and automatically loads them when ConsumerCheck is launched. When launching ConsumerCheck for the very first time no data are imported. One can import data by selecting *File -> Add Data sets* from the menu at the top of the GUI. When importing data for the first time after launching ConsumerCheck, regardless of whether data are already imported or not, a window appears providing short information on how each type of data should be structured. The information provided in this window is a short summary of what is described in section 2. Then a standard *Open file* dialog appears which allows for selection of one or more files for import. After clicking the *Open file* button in the open file dialog an import dialog appears for each selected file in successive order. The look and type of the import dialog depends on the format of the selected file and as such provides different parameter settings for the import. Note that only standard ASCII characters are allowed in the data file names as of PanelCheck version 1.1.0. Below a short description of the import settings for text and Excel files is given.

Data import dialog for text files

Fig. 7 shows a screenshot of the data import dialog for text files. At the top of the widget the path to the location of the file is displayed. Below is a grid providing a preview of the raw data that is about to be imported. This may be useful for quick inspection of whether the correct data were selected for import. The next import parameter provides a drop down menu where the encoding of the data may be selected. By default *ASCII* encoding is selected which fine to use if the data files do not contain special characters. The other two choices are *UTF-8* and *latin-1*. For more information on which encoding should be used, please consult (http://www.unicode.org/). Below there are three so-called radio buttons where the user can communicate to ConsumerCheck how each column of data in the text file is separated from one another, that is by 'tab', 'comma' or 'space'. Next, the user can set whether floats (that is numbers with decimals) are defined by commas or periods. Below, users can provide a name for the data set that will be used throughout ConsumerCheck. If the user doesn't give the data set a new name at this point, it is still possible to do so in the *Datas set* tab, that is described in section 6.2. Beneath there is a drop down menu where the user can define of which type the data set is. The five possible selections in the drop down menu are *consumer liking*, *consumer characteristics*, *product design*, *descriptive analysis / sensory profiling* and *other*. Those are the data types discussed from section 2.1 through 2.4 and 2.7. Selection of data type is important in order to have ConsumerCheck recognise the right data for the right statistical method. If the user does not set the data type in the import dialog it is still possible to do so later in the *Data sets* tab (see section 6.2). Eventually, the user can check or uncheck two check boxes to indicate whether the data have product and variable names included or not.

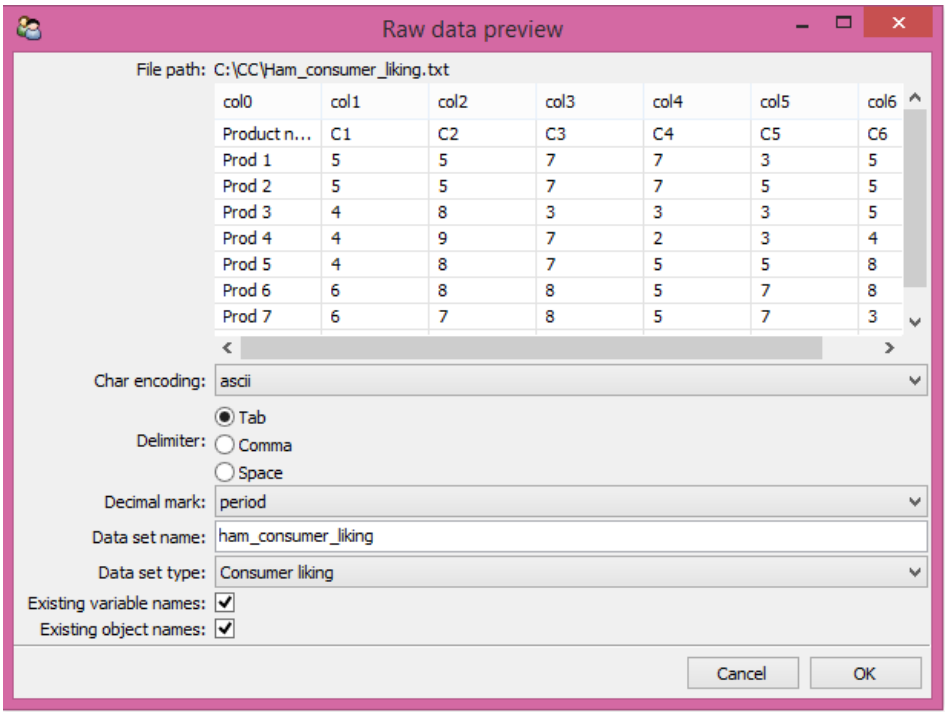


Figure 7: This is a screenshot of the import dialog for data stored in plain files such as **.txt** or **.csv**.

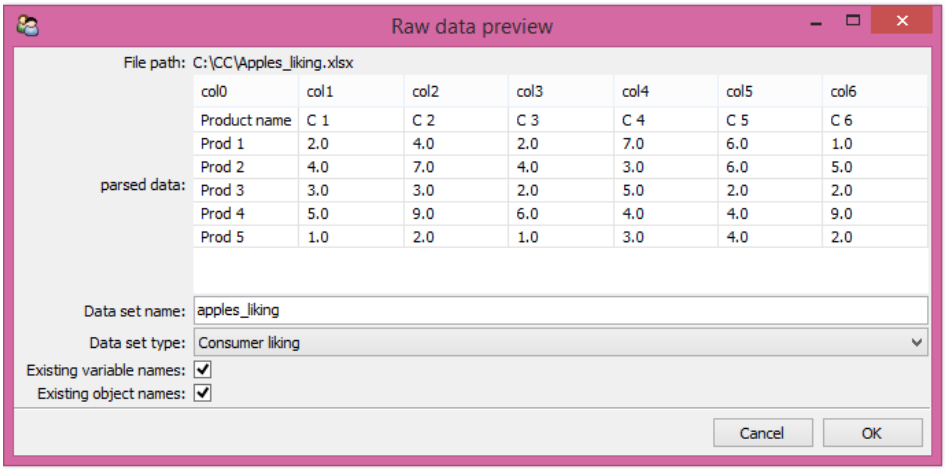


Figure 8: This is a screenshot that shows the import dialog for data stored in Excel files.

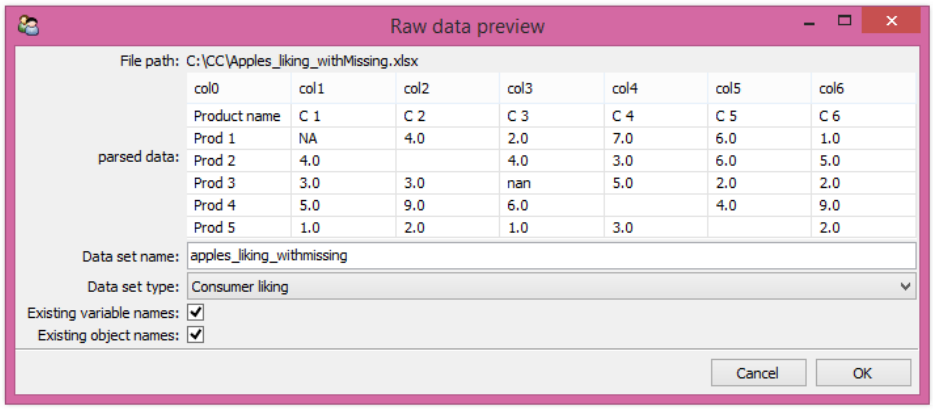


Figure 9: This is a screenshot of the Excel import dialog of data with missing values. Note that there are missing values in: col1-row2 marked as **NA**; col2-row3, col4-row5 and col5-row6 as *empty cells*; col3-row4 marked as **nan**. All three are valid approaches to mark missing values in the data.

Data import dialog for Excel files

Fig. 8 shows the dialog for data import from Excel files. Its structure and usage are almost identical to the import dialog for text files, except for that there are no settings for text encoding and options for delimiter. Both are detected automatically by ConsumerCheck.

Missing values in data

As mentioned earlier (see section 2) the import of data with missing values is allowed. There are plans to implement imputation routines for the handling missing values in future versions of ConsumerCheck, but progress will greatly depend on availability of funding and resources in general. Until then, if the tabs *Basic stat liking* (section 3.1 and 6.3), *PCA* (section 3.2 and 6.4) and *Preference mapping* (section 3.3 and 6.5), *PLSR/PCR* (section 3.4, 3.5 and 6.6) are to be used and the data contain missing values the users must impute missing values with their preferred imputation method outside ConsumerCheck prior to import into the software. Currently, the *Conjoint* method is the only method that handles missing values and returns computation results. When data with missing values are to be imported the missing values may be indicated either by leaving their respective cells either empty or by marking them as **NA** or **nan**. Fig. 9 shows an example of importing data with missing values.

Removing data

If some data set needs to be removed from ConsumerCheck this can be done easily by selecting the following from the main menu: *File -> Remove Data sets*.

6.2. Data sets tab

The *Data sets* tab is the first of several tabs of the GUI. Its main purpose is to provide a general overview of the data imported, a short summary of each data, tools for setting data parameters and methods for processing data. The *Data sets* tab, like all other tabs in the GUI, is divided into a left and right panel. Fig. 6 shows and example of what the the *Data*

sets tab looks like with five data set imported.

The left panel shows a so-called tree-control with one data set at each branch. With a single left-click on a data set information that is specific for this data set is shown in the right panel, i.e. the *data set name*, *data set type* and *data set summary*. In the *dat set name* text field at the top of the right panel the name of the particular data set may be changed. The name defined in this text field is then used in the statistical method specific tabs in ConsumerCheck (see sections 6.3 to 6.7). With the *data set type* drop-down menu right below the type of the data set may be set. Both the name and the type of the data may have been set already in their respective import dialogs (see section 6.1), but here the user can change these parameters again if needed. Below, the *data set summary* provides a short summary of the respective data, such as the dimension of the data, the mean and standard deviation across all entries as well as the minimum and maximum values in the data. A double left-click on a data set in the tree-control generates a new window that visualises the data in a sheet. From that window one can copy the data by clicking on the *copy to clipboard* button and paste it into other software applications such as Excel or Open Office Calc Spreadsheet. A single right-click on a data set invokes a menu that lets the user do various things with the data. At the time of writing, this menu contains two options, but more may follow in future versions of ConsumerCheck: (I) *Create transposed copy* and (II) *Delete*. The first option allows the user to make a transposed copy of the selected data set, meaning that a copy of that specific data is added at the lower end of the tree-control, but where rows have become columns and columns have become rows. This may be useful when applying the multivariate statistical regression methods *PLSR* (see section 3.4) and *PCR* (see section 3.5) to two data sets as it is done with the *PLSR/PCR* tab (see section 6.6). The second option lets the user delete data sets from ConsumerCheck. The data then are no longer available at the tree-control. If needed again, the data may be imported the usual way as described in section 6.1.

6.3. Basic stat liking tab

The purpose of *Basic stat liking* tab (see screenshot in Fig. 10) is to provide visualisation and simple analysis of consumer liking data to the user. This implies that only data of type *consumer liking* are listed in the *Select data set* box in the upper right corner of the GUI and as such are available for visualisation and analysis. As seen in the Fig. 10, in this case consumer liking data from the apple and ham data set are present and available for visualisation. At the left there is a tree-control from which plots may be generated by double left-clicking on tree-control items. The tree-control is dynamic and expands or retracts as consumer liking data are checked or unchecked. The tree-ctrl provides three types of plots: *box plots*, *stacked histogram plots* and *single product histogram plots*. Each type will be described below.

Plots for all products - Box plot

Fig. 11 shows the box plot for the *ham consumer liking* data where consumers rated 8 food products on a hedonic scale from 1 to 9, where 1 represents "don't like at all" and 9 represent "like very much". The box plot shows that across all consumers each product received the highest (9) and lowest (1) rate by at least one consumer. This is visualised by the vertical lines that extend from 1 to 9 for each product. The green boxes for each line visualise the distribution of the ratings between the 25th and 75th percentile. The dark green line across the green boxes shows the median rating for that product. Note that the plot can be saved

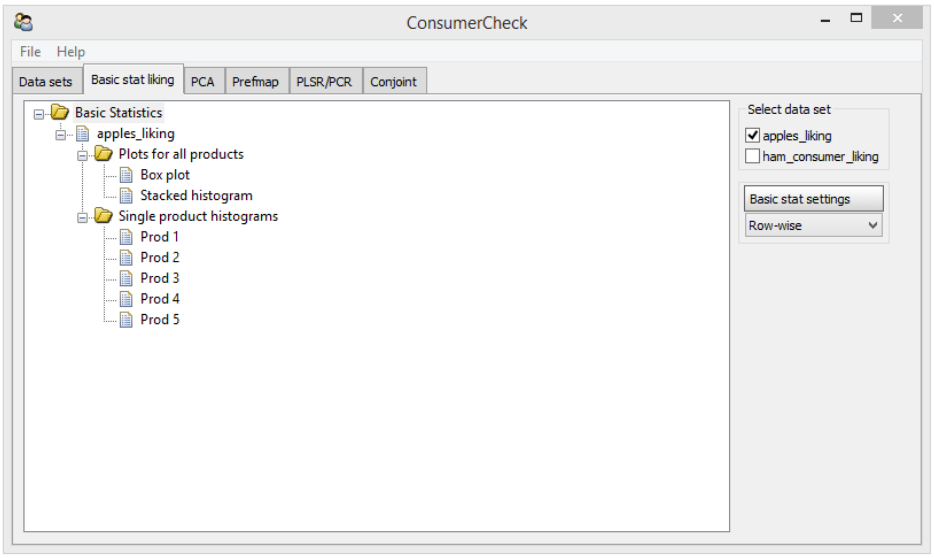


Figure 10: This screenshot shows the graphical user interface of the *Basic stat liking* tab.

in .png format when clicking on the photo camera icon placed in the lower left corner of the plot window. This is a common feature for all plots implemented in ConsumerCheck.

Plots for all products - stacked histograms

Stacked histograms provide another and richer way of visualising *consumer liking* data. In Fig. 12 a stacked histogram plot is shown for the same data as presented earlier in a box plot in Fig. 11. Along the horizontal axis again the products are shown, while the vertical axis displays either the number of consumers or a percentage of the total number of consumers. If percentages are to be shown, the *Percent* checkbox at the bottom of the window needs to be checked. With the stacked histogram each bar represents one product and each colour in the bar represent a certain rating of the product. For *prod 3* one can see that 14 consumers or 17% of the total number of consumers rated this product with 1 (”don’t like at all”). 5 consumers or 6% of the consumers rated *prod 3* with 2, and so on. In this way the distribution of the ratings is visualised in a more detailed way than in the box plots.

Single product histograms

Single product histograms show for each product the distribution of the liking ratings in separate histograms. Fig. 13 shows an example for *prod 1*.

Now, instead of putting all information into one bar as seen in the stacked histogram plot, one plot is dedicated to *prod 1* alone. In the single product histogram the bars represent increasing liking from left to right. For each rating the percentage of consumers having rated the product this way is displayed on the top of the bar.

Column-wise summary of consumer liking data

The main idea behind the *Basic stat liking* tab is to provide information on the distribution of liking ratings with focus on the products / objects in the *consumer liking* data. It is,

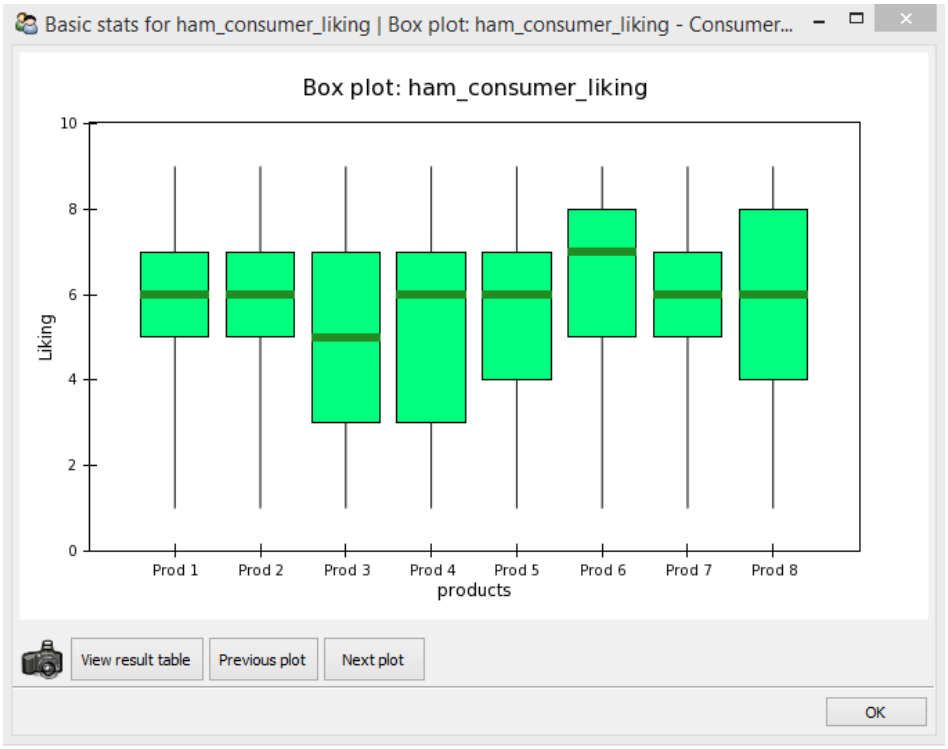


Figure 11: This is an example of what the box plot looks like for the *ham consumer liking* data. The plot is found under the tab named *Basic stats liking*

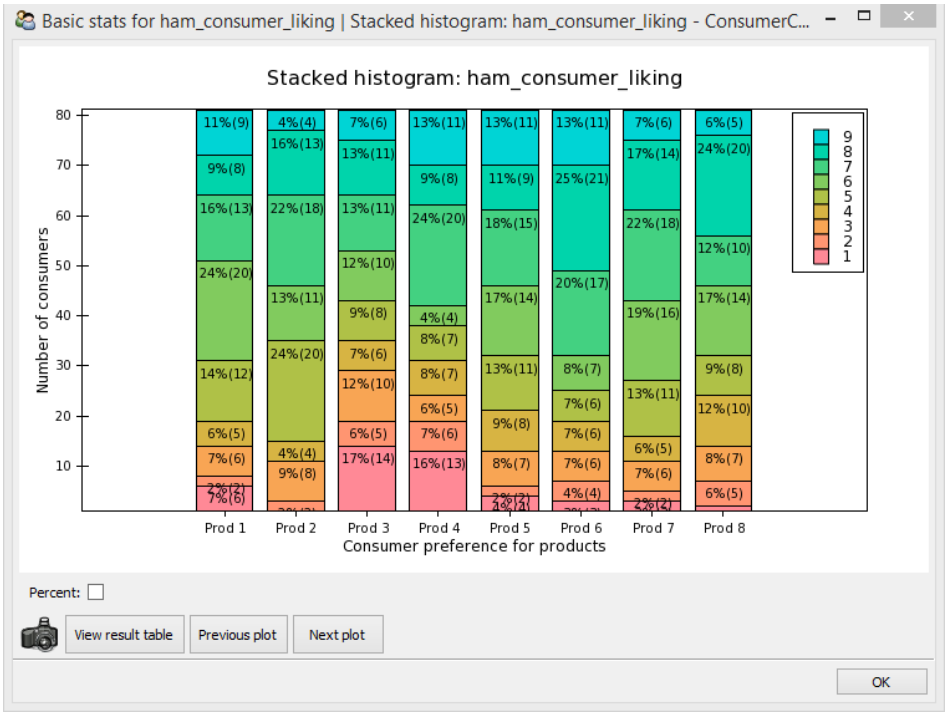


Figure 12: This is an example of what a stacked histogram *ham consumer liking* data. The plot is found under the *Basic stat liking* tab.

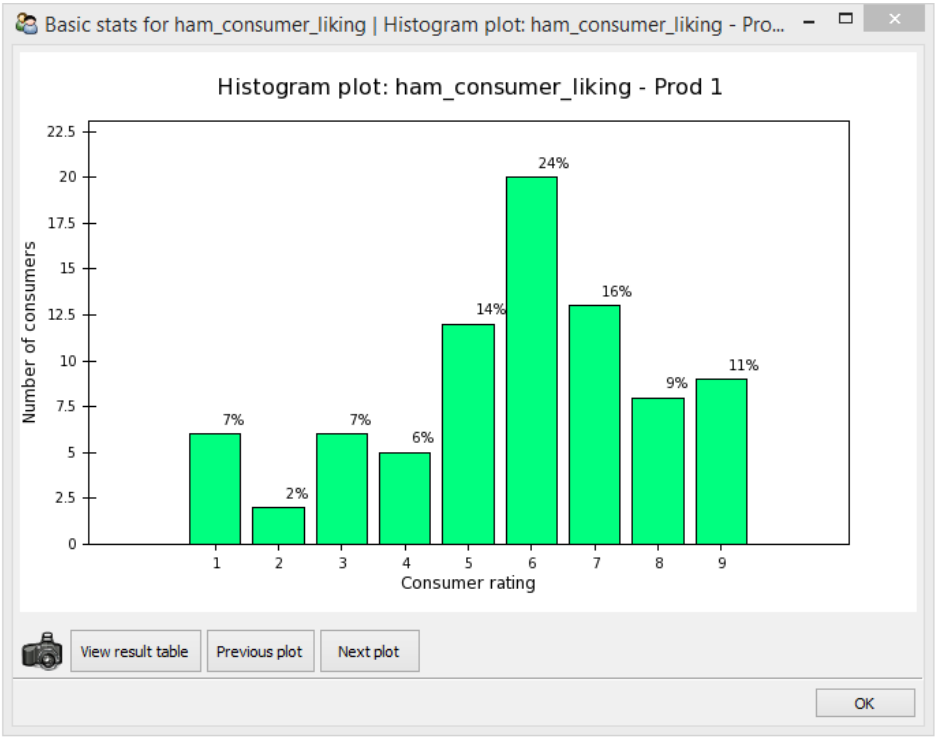


Figure 13: This screenshot shows the histogram for product *prod 1* of the *ham consumer liking* data.

however, possible to visualise the liking distributions also with focus on the consumer. This can be achieved by selecting *Column-wise* in the *Basic stat settings* drop-down menu at the right side of the GUI. Double clicking on *Box plot* and *stacked histogram* in the tree-control now generates box plots and stacked histograms, respectively, for each consumer. Note that the default for the *Basic stat setting* is *row-wise*, i.e. focus on the products.

6.4. PCA

When selecting the PCA tab the user can carry out principal component analysis on data of type *descriptive analysis / sensory profiling*, *consumer liking* and *consumer characteristics*. Fig. 14 shows an example of the PCA tab with four of data sets ready for analysis. Throughout the PCA section the *apple descriptive analysis / sensory profiling* data (as described in section 2.6) will be used to illustrate what results are provided when analysing data with PCA. Again, the tree-control for generating plots is on the left side and the data available for analysis with PCA are listed in the upper right corner under *Select data set*. Below, the user can choose more settings for the computation of the PCA model. By checking the *Standardise* checkbox, all variables in the data are standardised such that they have zero mean and a standard deviation that equals one. By default, that is when the *Standardise* checkbox is unchecked, variables are mean centered. Note that variables with zero variance across objects/rows are left out of analysis when the *Standardise* checkbox is checked. This is because variables with zero variance cannot be standardised (division by zero). In such a case a message box will inform the user about leaving out such a variable. Moreover, the user

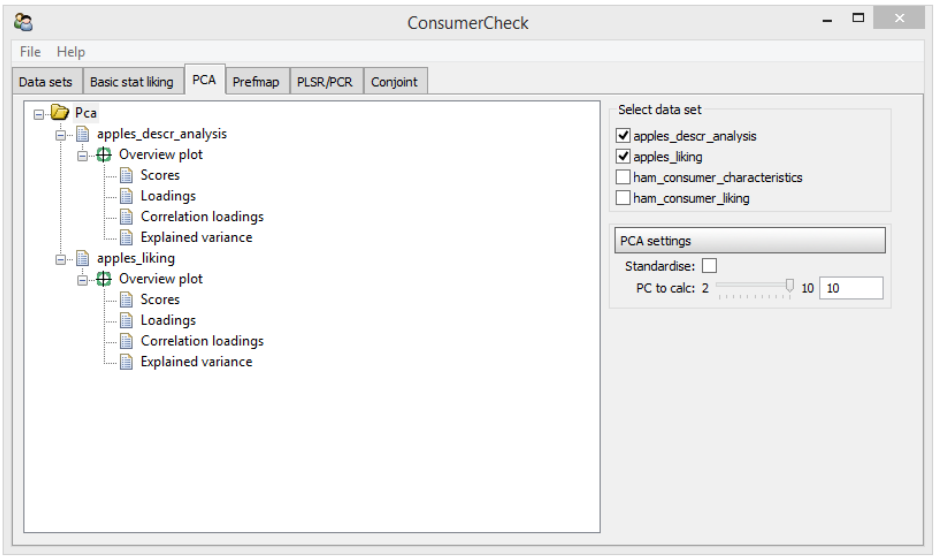


Figure 14: A screenshot of the *PCA* tab.

can select how many principal components (PC's) are to be computed for the PCA model. This can be done either by typing the number of wanted PC's directly into the text box or by using the slider at its left side. Note that the maximum number of PC's that can be computed from a data set is equal to either *number of variables* or *number of objects* in the data set, whichever is smaller. Below, we will discuss which plots may be generated from the tree control.

PCA - Overview plot

By double-clicking on the tree control item named *PCA overview plot* a new window will appear that consists of four sub-plots. An example of such an overview plot is shown in Fig. 15. As can be seen the four sub-plots are PCA *Scores*, *Loadings*, *Correlation loadings* and *Explained variance*. More details on each plot are given below in the respective subsections. With a single left-click on one of the four subplots a new window will appear showing an enlarged version of that specific plot. The same can be achieved by directly double left-clicking on the respective item in the tree control.

PCA scores

The PCA scores plot visualises how the objects or products from the analysed data matrix are distributed across the space spanned by two principal components (PC). By default the plot shows the scores for PC1 and PC2, that is the components explaining the highest and next to highest variance in the data. Fig. 16 shows a example of what a PCA scores plot may look like. Here PC1 and PC2 explain 90% and 9% of the calibrated variance in the data, respectively, totalling 99%. In other words, almost all of systematic variation in the data is visualised by these two components. From the plot we can see that product 3 and 5 are very similar since they are located very close to each other. At the other side of the plot there are products 1, 2 and 4 indicating that these products are very different from product 3 and 5 given the fact that PC1 explains 90% of the calibrated explained variance. Product 1 and 2

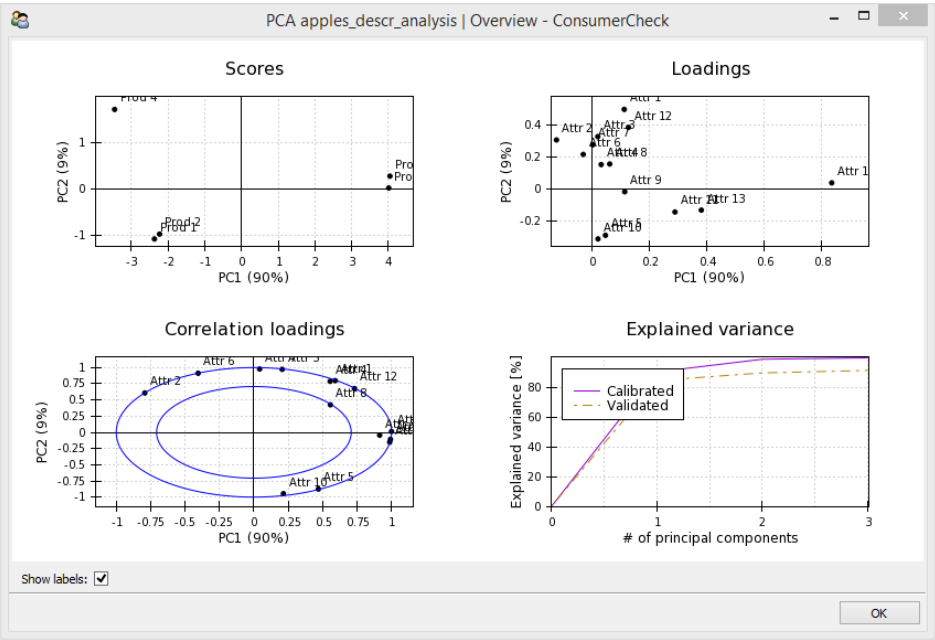


Figure 15: Screenshot of an *Overview plot* for the PCA model based on the *apple descriptive analysis / sensory profiling* data.

are very similar because of their proximity in the plot. Product 4 is not too different from product 1 and 2 with regard to PC1, but some differences are present since the products are spread out across PC2 which accounts for 9% of the variance in the data.

PCA loadings

The PCA loadings plot visualises how the variables, which in our case are sensory attributes describing the food product, contribute to the variation in the data. Fig. 17 shows the PCA loadings for the data. Clearly, attribute 14 contributes much to the variation explained by PC1 since it has a large absolute loading for PC1 compared to all other sensory attributes. Attribute 11 and 13 are two other variables contributing much to the variation explained by PC1. For PC2, attributes 5 and 10 on one side and attributes 1, 2, 3, 7 and 12 on the opposite side are variables contributing most to variation. In general, variables that are located close to each other are highly correlated to one another with respect to the plotted principal components and vice versa. The closer a variable to the origo, the less it contributes to systematic variation explained by the two visualised PC's.

By superimposing the PCA scores and loadings plot one can get more information on the products. Both attribute 14 and products 3 and 5 are located on the right side of the loadings and scores plot, respectively. This means that these two products have high values for attribute 14 while products 1, 2 and 4 have lower values for attribute 14, since they are located on the opposite side with regard to PC1. Note that in this example the data are not standardised, which is why attribute 14 is dominating. The reason for not standardising the variables here is that the trained sensory panel scores all attributes on the same scale, which in our case is from 1 (low intensity) to 9 (high intensity). Elaborating this data further we can

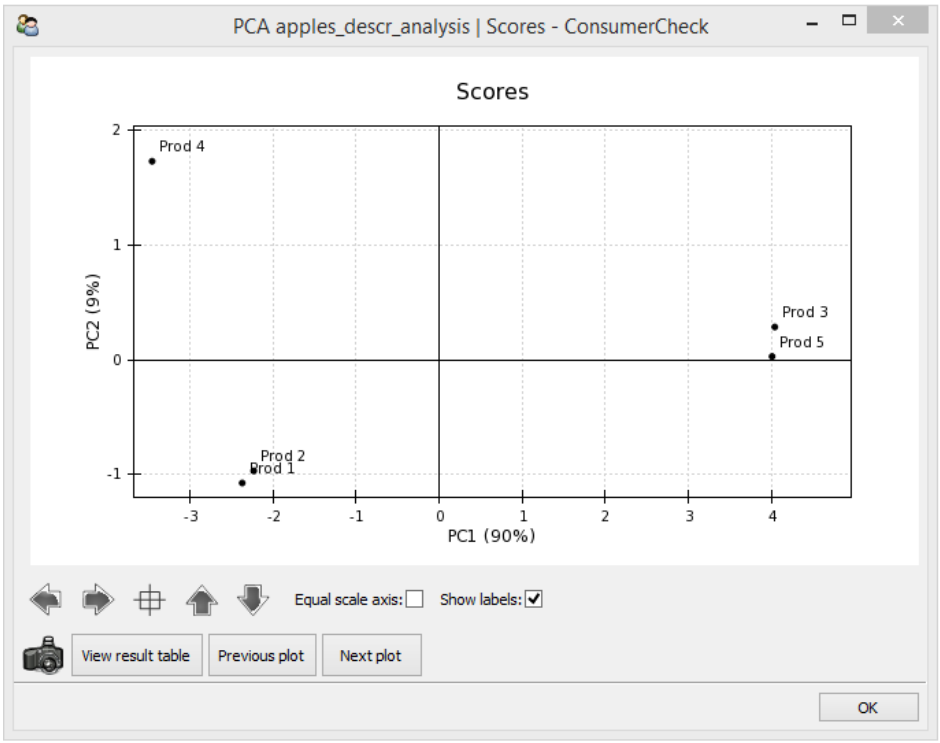


Figure 16: Screenshot of a PCA scores plot showing how the five products from the *apple descriptive analysis / sensory profiling* data are distributed across PC1 and PC2. PC1 and PC2 together describe 99% of the variation in the data.

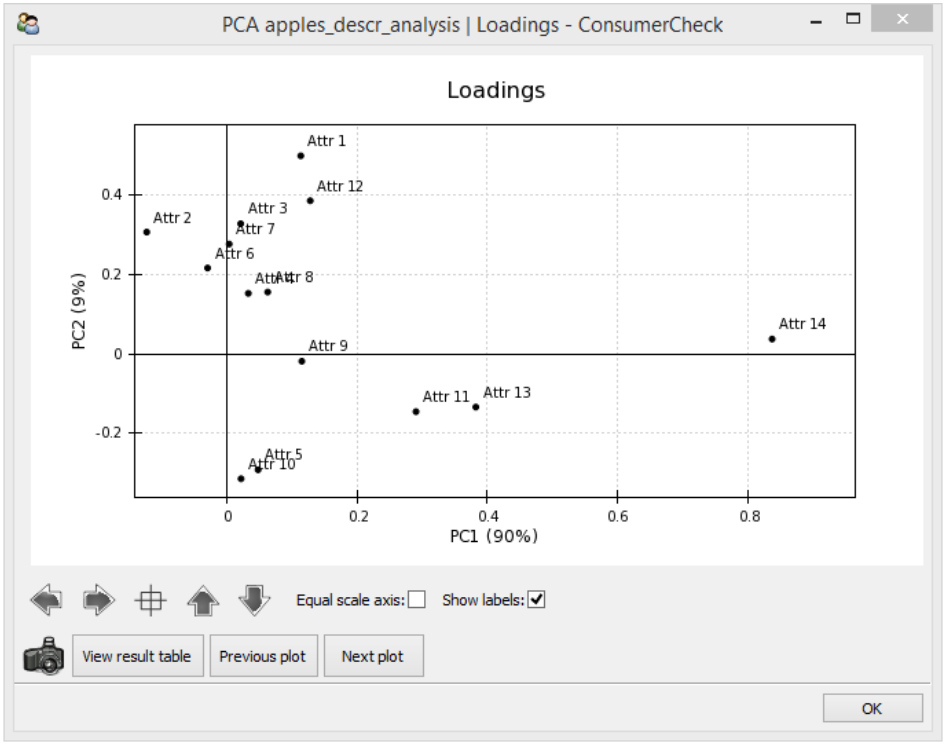


Figure 17: Screenshot of a PCA loadings plot for the *apple descriptive analysis / sensory profiling* data.

see that product 1 and 2 have high intensities for attribute 5 and 10 and lower intensities of attribute 1 and 12 (which are located at the opposite side with regard to PC2). For product 4 the opposite is true.

PCA correlation loadings

PCA correlation loadings, as shown in Fig. 18, are another way of visualising the contribution of the variables to the total variance in the data. PCA correlation loadings provide information on how systematic the variance of a variable is with regard to the computed PC's, not only how much variance was contributed by the variable (as visualised in the PCA scores plot). More precisely, a correlation loading is actually the correlation between the original data of a specific variable and the scores of a specific PC (NEED REF). In this way one can see to which degree the variation from a specific variable is systematic or rather noisy, regardless of the total variance it contributes. The two rings in the correlation loadings plot in Fig. 18 indicate specific amounts of explained variance for the attributes at hand. The outer ring represents 100% explained variance while the inner ring represents 50% explained variance.

Consider an example with attributes 4 and 8. When looking at the loadings plot in Fig. 17 these two variables are located close to each other contributing about the same amount of variance to the variance explained by PC1 and PC2. In the correlation loadings plot (Fig. 18), however, they are no longer located close to each other. Attribute 8 is located just inside the inner ring, which indicates that just under 50% of the variation of this variable is explained by PC1 and PC2. Remember that that PC1 and PC2 together explain 99% of the total variance

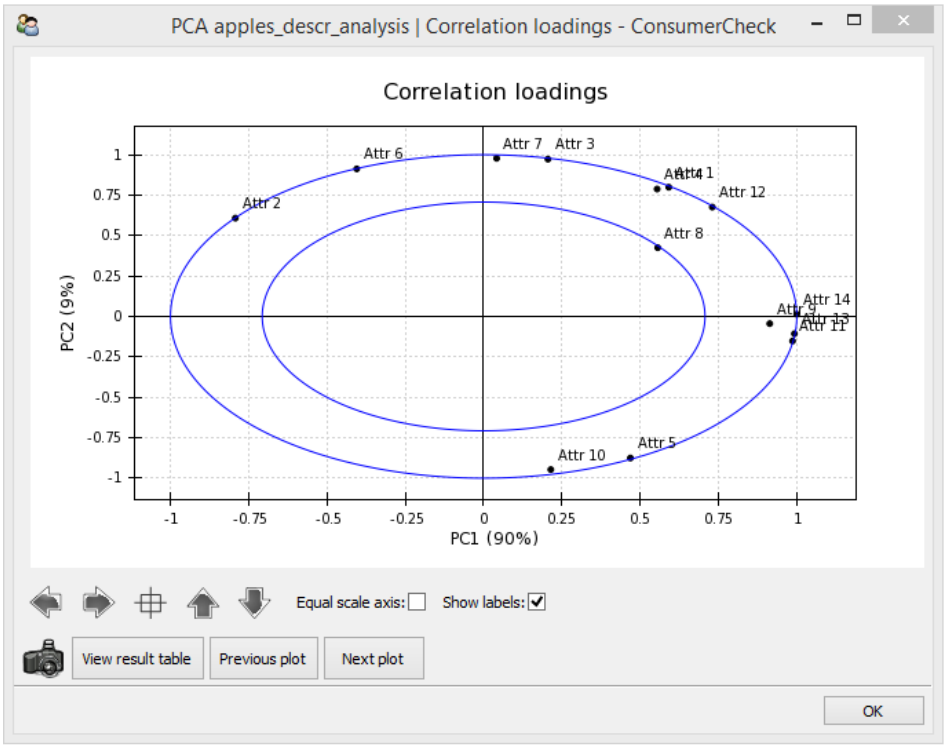


Figure 18: PCA correlation loadings for the *apple descriptive analysis / sensory profiling* data.

in the data and that remaining higher PC's provide very little or no information. This means that that not much more than 50% of the variance of attribute 8 will be explained by the higher PC's and that the remaining variance of that variable is likely to be noise. Attribute 4 is very close to the outer ring, meaning that almost 100% of its variation is explained by PC1 and PC2, thus indicating that its variance is very systematic. In this way it is possible to see that the variation of attribute 4 is much more systematic with the variance described by PC1 and PC2 than that of attribute 8, even though both of them contribute about the same amount of variance to the data. Considering this, PCA correlation loadings are a useful complement to the PCA loadings for better understanding of how variables contribute to the total variance in the data.

PCA explained variances

Fig. 19 shows the calibrated and validated explained variances of the same data. It can be seen easily that with only two PC's almost all of the variance in the data is explained by the model (99% as mentioned above). For model validation full cross-validation (also known as leave-one-out) was applied to the data. The resulting validated explained variance rises to about 85% with PC1 and then slightly increases to 90% after PC2, thus confirming that the model is robust.

6.5. Preference mapping

Preference mapping is applied simultaneously to *consumer liking* data and *descriptive anal-*

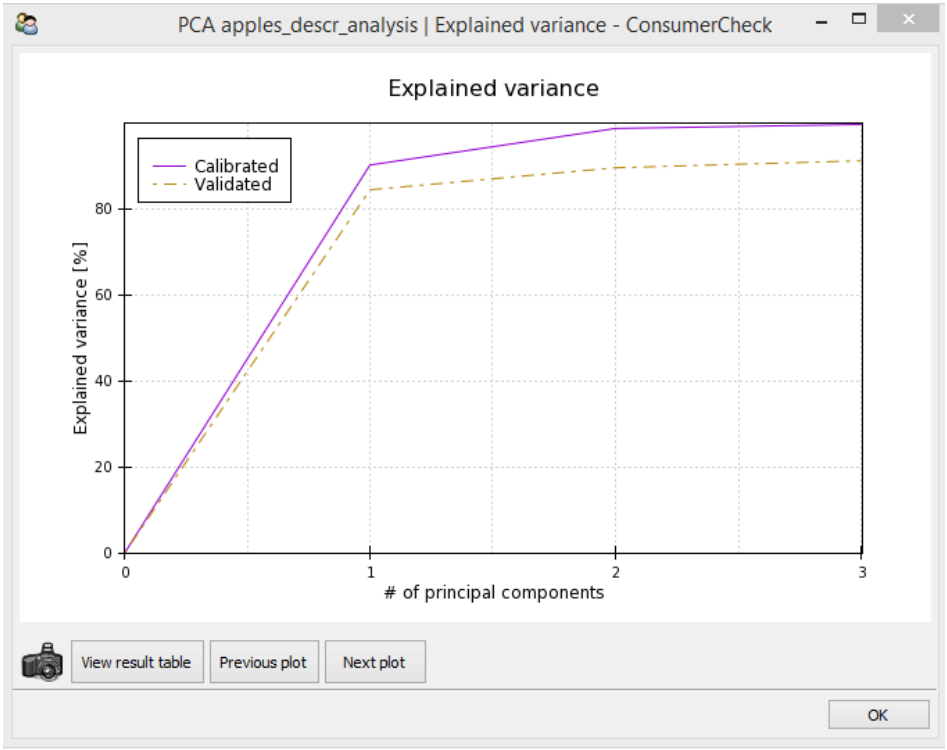


Figure 19: Calibrated and validated explained variance of the PCA model for the *apple descriptive analysis / sensory profiling data*.

ysis / sensory profiling data. The aim is to find drivers of liking that may determine why some products are preferred over other. The two standard statistical tools applied to build a preference mapping model are partial least squares regression (PLSR) and principal component regression (PCR). Both are implemented in ConsumerCheck and are coded in Python.

When building a preference mapping model, both consumers and the trained sensory panel need to evaluate the same set of products. In each data, the row order of the products needs to be identical otherwise wrong data are linked together and results will lead to incorrect conclusions. Two drop down menus on the right side of the GUI let the user define which data are to be linked (see Fig. 20). The left drop down menu contains all data imported into ConsumerCheck that were tagged as *consumer liking* data whereas the right drop down menu contains all data tagged as *descriptive analysis / sensory profiling* data. If the number of rows in the two data set do not match because they originate from different experiments, ConsumerCheck will give an error message. Once two matching data set are selected a tree control appears on the left side of the GUI. There are a few more settings for computation of the model at the right side of the GUI that can be adjusted by the user. First, the user may choose between internal preference mapping (*consumer liking* data are set as the X matrix in the model and the *descriptive analysis / sensory profiling* data are set as the Y matrix) or external preference mapping (*descriptive analysis / sensory profiling* data are set as X and *consumer liking* data are set as Y). With the next setting the user may choose between the statistical methods PLSR or PCR. Below there are two check boxes where the user may choose to standardise either or both X and Y. By checking the *Standardise* checkboxes, all variables in the data are standardised such that they have zero mean and a standard deviation that equals one. By default, that is when the *Standardise* checkboxes are unchecked, variables are mean centered. Note that variables with zero variance across objects/rows are left out of analysis when the respective *Standardise* checkbox is checked. This is because variables with zero variance cannot be standardised (division by zero). In such a case a message box will inform the user about leaving out such a variable. With the last parameter setting the user can select the number of components to be computed for the model.

Preference mapping - overview plot

By left double-clicking on *Overview plot* a new window opens that shows the X scores (upper left), X & Y correlation loadings (upper right), explained variance in X (lower left) and explained variance in Y (lower right) as shown in Fig. 21. The respective plots can be accessed directly by left-double clicking on their respective item in the tree control or by a single left-click directly on the overview plot. More information on each plot is provided below. Furthermore, from the tree control separate plots for X and Y correlation loadings may be generated.

Preference mapping - X scores

The results presented in this section were computed with the following settings: internal preference mapping (i.e. the *consumer liking* data are set to be X in the model and *descriptive analysis / sensory profiling* data are set to be Y); PLSR; X and Y are not standardised since all variables in the respective matrices are based on the same scale. Fig. 22 shows the X scores of the preference mapping model visualising how the products relate to each other in the space

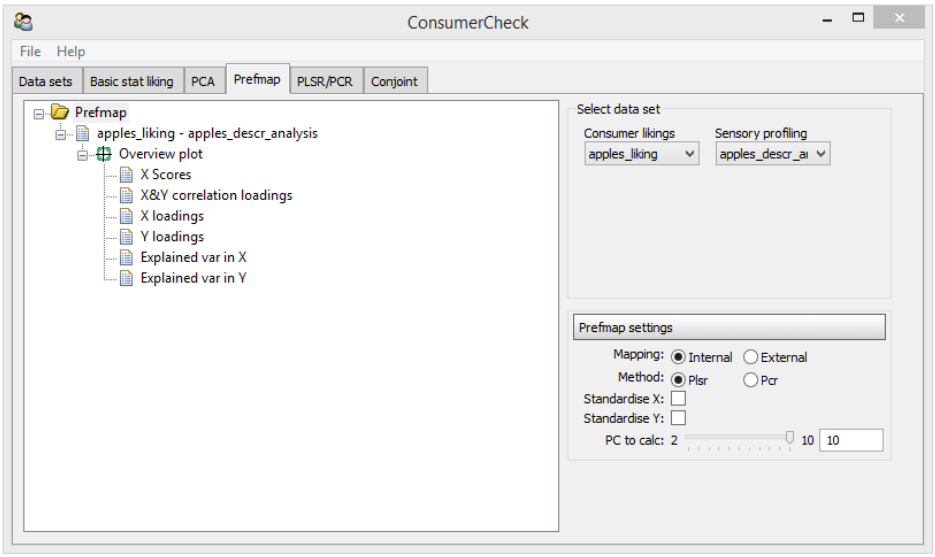


Figure 20: Screenshot of the GUI for *preference mapping*. The tree control to the left is shows which plots may be generated for the preference mapping model based on the *apple consumer liking* data and *apple descriptive analysis / sensory profiling data*.

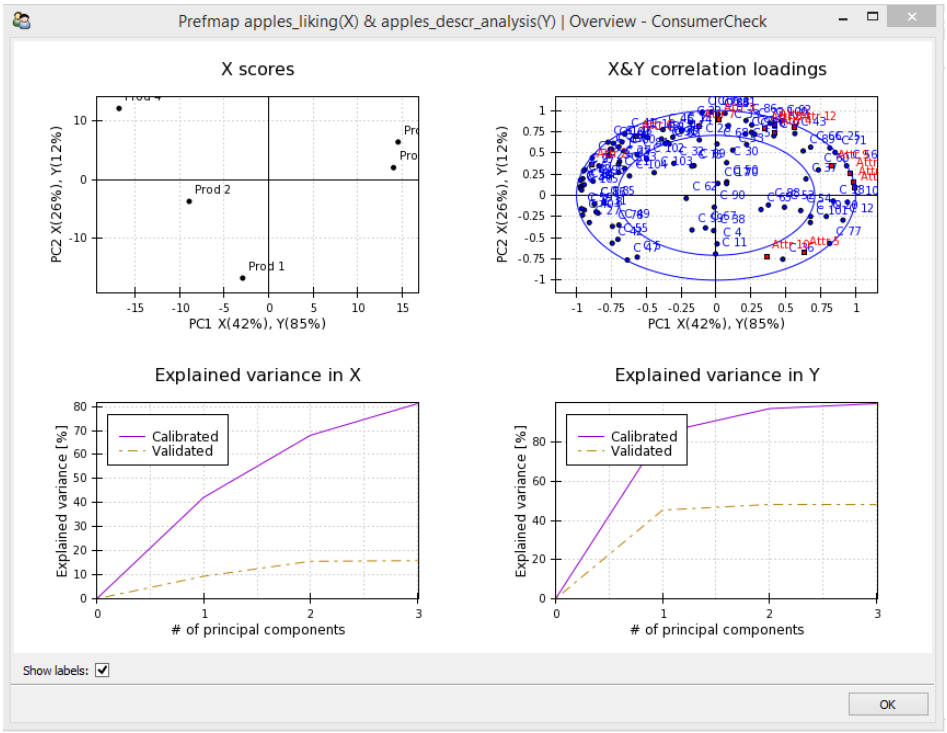


Figure 21: Overview plot for preference mapping model for the *apple consumer liking* data and *apple descriptive analysis / sensory profiling* data.

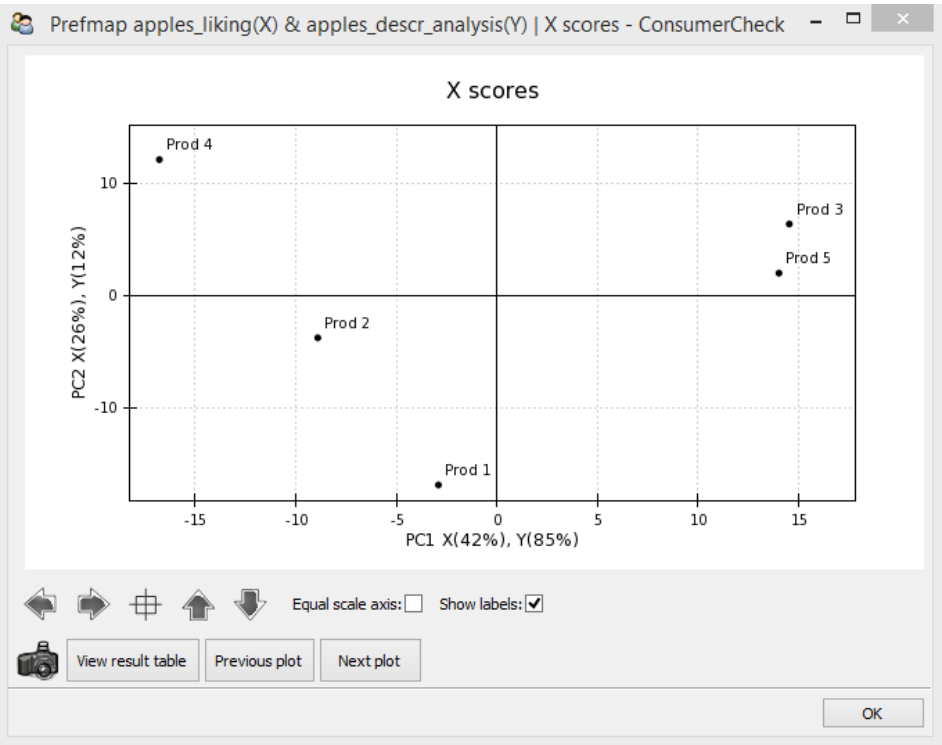


Figure 22: A screenshot showing the X scores plot in preference mapping for the *apple consumer liking* and *descriptive analysis / sensory profiling data*.

spanned by the first two components. As with the PCA scores plot in Fig. 16 similar products are located close to each other and dissimilar products have a larger distance between them. This time, however, the distribution of the products is influenced by the common variance in both X and Y matrix, since in this case PLSR was chosen to compute the preference model. Remember that if PCR is chosen for model computation instead of PLSR, the X scores are computed from matrix X only, while matrix Y has no influence. For details on differences between PLSR and PCR the reader is suggested to consult section 3.4 and 3.5.
As can be seen from Fig. 22 product 3 and 5 are again located close to one another. Products 1, 2 and 4 are again on the opposite side of product 3 and 5 with regard to PC1, however, they are more scattered than they were as with PCA where analysis was based on *descriptive analysis / sensory profiling* data only (section 6.4). It is important to note how much of the variance in X and Y the first two principal components explain. PC1 and PC2 explain 42% and 26% (first number in parenthesis) of the variance in the X matrix. This totals to 68% for X which is considerable taking into account how noisy *consumer liking* data often can be. PC1 and PC2 explain 85% and 12% (second number in parenthesis) of the data in Y (the *descriptive analysis / sensory profiling* data in our case) totalling 97%. The high levels of explained variance for X and Y indicate that there is a lot of common systematic variation in the data.

Preference mapping - X&Y correlation loadings

Fig. 23 shows the actual preference map that is used for interpretation and visualisation of

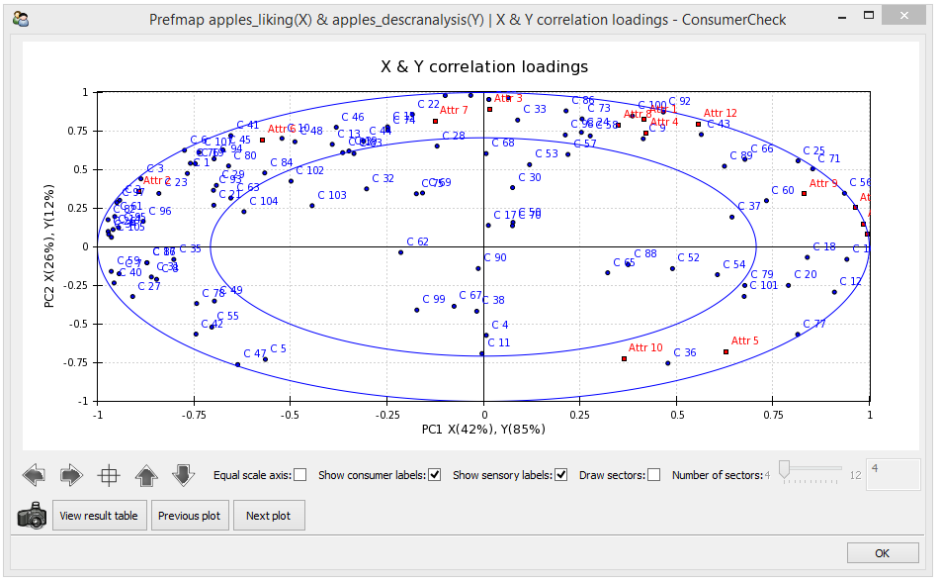


Figure 23: A screenshot of the X & Y correlation loadings plot, the actual preference map for the *apple consumer liking* and *descriptive analysis / sensory profiling data*.

consumer preferences and drivers of liking. In this plot both the correlation loadings from X and Y are displayed in the same plot.

Correlation loadings belonging to matrix X are always coloured in blue. In this example they start with the letter 'C' followed by a number that identifies the consumers that participated in the trial. The correlation loadings from matrix Y are always coloured in red. In this case they are the sensory attributes that describe the product. What we can conclude from Fig. 23 is that many consumers prefer products with high intensities of attribute 2 and 6 (upper left part of the plot) since a large part of the consumers are in proximity of those attributes. Attributes 11, 13 and 14, which are all highly correlated, are less preferred although there are a few consumers that prefer high intensities of these sensory attributes. All of them have high explained variances, since they are located very close to the outer ring that indicates 100% explained variance. Attributes 5 and 10 are also correlated, however to a lesser degree. The explained variances for those two attributes are somewhat lower. Remember that the inner ring indicates 50% explained variance. Consumers in the inner circle closer to the origo don't discriminate between the products with regard to the variation described by PC1 and PC2. Since the X & Y correlation loadings plot often is crowded it may be helpful to remove the consumer or sensory attribute labels in order to get a less distorted picture of where consumers and products are located in the plot. This can be done by checking/unchecking the respective boxes ("Show consumer labels", "Show sensory attribute labels") at the bottom of the plot. Furthermore, the X & Y correlation loadings plot can be divided into segments when checking the "Draw sectors" checkbox. This may be a handy tool to identify quickly which products and attributes are most preferred, that is which products and attributes have most consumers in their proximity in the plot (when correlation loadings plot and X scores are superimposed). By default four segments are drawn as shown in Fig. 24. The number of segments may be changed by either moving the slider located to the right of the checkbox or by entering the number of segments in the text box to the far right.

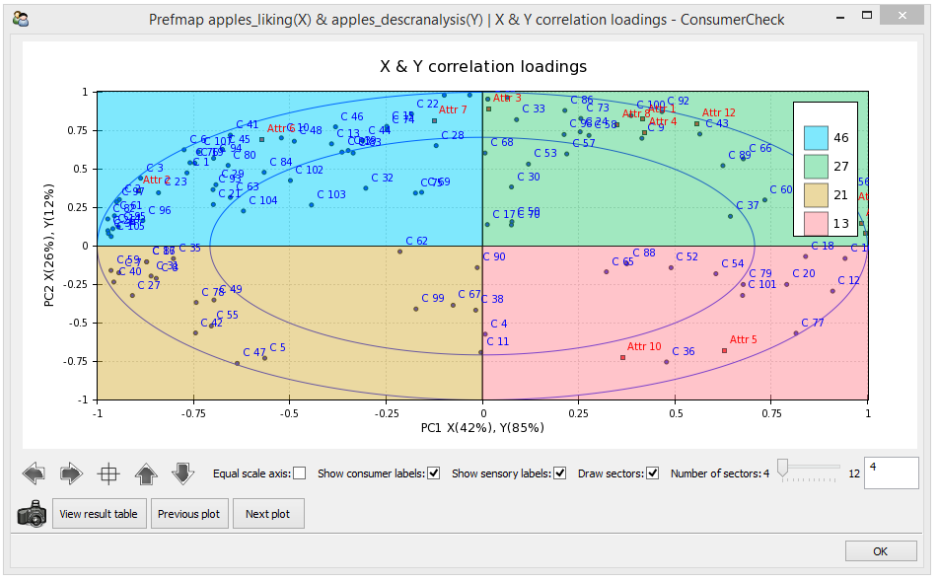


Figure 24: A screenshot of the X & Y correlation loadings plot, the actual preference map for the *apple consumer liking* and *descriptive analysis / sensory profiling data*. This is the same plot as shown in Fig. 23, however here the X & Y correlation loadings are devided into 4 segments.

As can be seen most consumers are found in the upper left segment which is coloured in blue. The legend indicates that the number of consumers in this segment is 46. The segment with the fewest consumers is found in the lower right corner coloured in pink containing only 13 consumers.

Preference mapping - X loadings

The X loadings in preference mapping show how the variables of the X matrix contribute to the common variation between X and Y for each principal component. Fig. 25 shows an example of X loadings for PC1 and PC2. As mentioned previously, *consumer liking* data were chosen to be the X matrix in the model, hence the variables of the X matrix are consumers that have tested the products. In Fig. 25 we can see how consumers spread out across the plane spanned by PC1 and PC2, providing information on how much variance every consumer contributed to the variance explained by PC1 and PC2.

Preference mapping - Y loadings

The Y loadings in preference mapping show how the variables of the Y matrix contribute to the common variation between X and Y for each principal component (TRUE FOR PLSR, NOT TRUE FOR PCR). Fig. 26 shows an example of Y loadings for PC1 and PC2 visualising how much variation each variable contributes to the explained variance described by PC1 and PC2.

Preference mapping - explained variances in X

Fig. 27 shows the calibrated and validated explained variances for the X matrix. One can

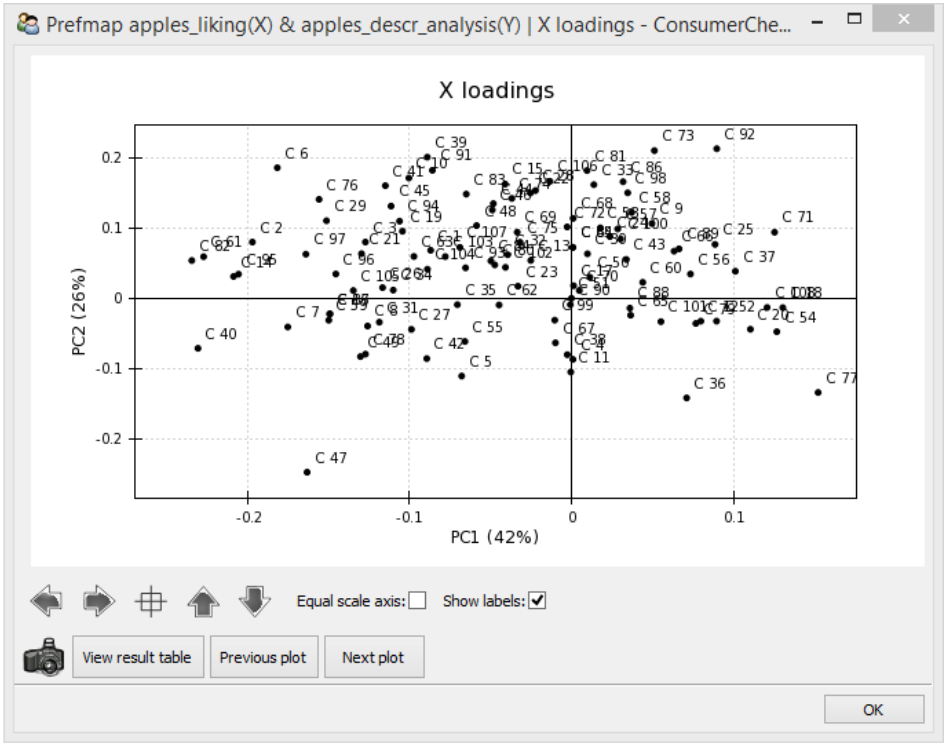


Figure 25: A screenshot of the X loadings in preference mapping based on the *apple consumer liking* data.

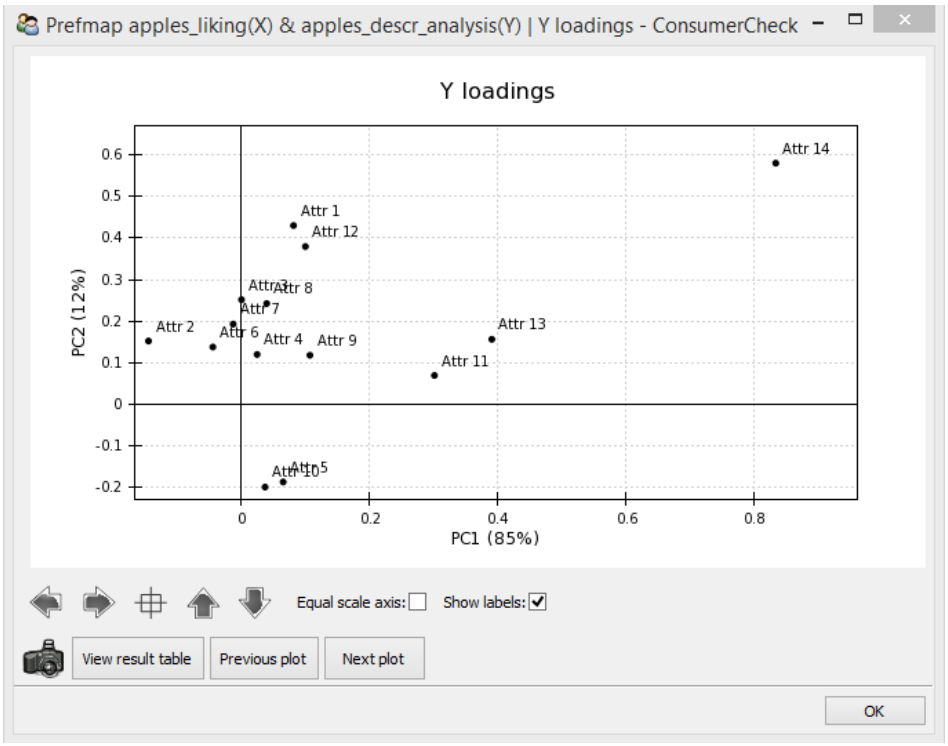


Figure 26: A screenshot of the Y loadings in preference mapping based on the *apple descriptive analysis / sensory profiling* data.

see that the calibrated explained variance increases to about 80% with the first three PC's (MAKE A NEW PLOT WITH MORE PCS). The validated explained variance, however, reaches only a level of about 17%. Low validated explained variances are quite common for *consumer liking* data since these data often are relatively noisy due to the individual differences between consumers and also because of the low number of objects or products in the data. When validating the model with cross validation, the model may change relatively much with each validation step which leads to poor predictions of the products left out in the cross validation process. The numerical results can be viewed by clicking on the *View result table* button which opens a new data window displaying the numbers. If needed, the user may select all or parts of the data and use the *copy to clipboard* button at the bottom of the data window to copy and paste the data to another software.

Preference mapping - explained variances in Y

Fig. 28 shows the calibrated and validated explained variances for the Y matrix. Here, the calibrated explained variance jumps up to about 85% and approaches 100% with the first two PC's. This is quite common for *descriptive analysis / sensory profiling* data, since trained panels typically produce more systematic data than untrained consumers. Numerical results can be accessed by clicking on the *View result table* button. A new data window appears then where numbers are viewed and may be copied to other softwares.

6.6. PLSR/PCR tab

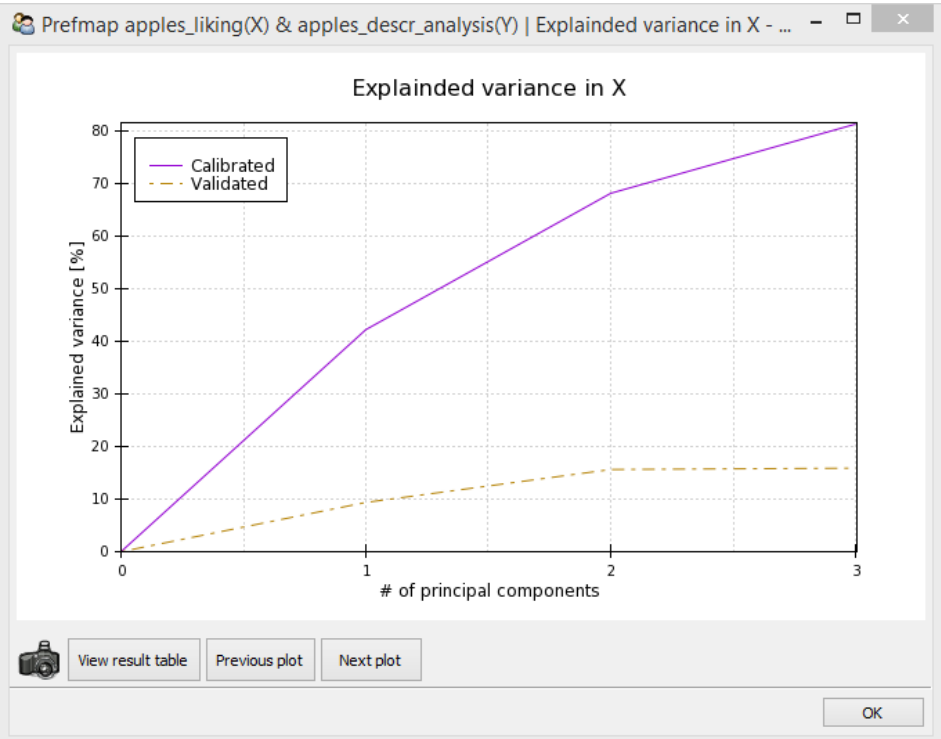


Figure 27: A screenshot of the explained variances in X for the *apple consumer liking* data in preference mapping model.

The implementation of the *PLSR/PCR* tab is actually almost identical to the *preference mapping* tab that was described above (see section 6.5). The most important difference is that the *PLSR/PCR* tab does not restrict its use to only *consumer liking* and *descriptive analysis / sensory profiling* data, but also allows for analysis of data tagged as *consumer characteristics*, *design matrix* and *other*. Another difference is that the *PLSR/PCR* tab doesn't provide the functionality for segmentation in the X & Y correlation loadings plot. Furthermore, there are some minor differences regarding the structure of the GUI when compared to the GUI of the *Prefmap* tab (compare Fig. 20 and 29). In the upper right corner of the *PLSR/PCR* tab the user can choose with the two drop down menus which data is set to be X and Y in the statistical model. Recall that in the *Prefmap* tab at the same place the user is supposed to set the *consumer liking* data and *descriptive analysis / sensory profiling* data for which the *preference mapping* model will be computed for. The other difference between the respective GUI's is that there are no so-called radio-buttons in the *PLSR/PCR* tab for the selection of *internal* or *external preference mapping*, since here the multivariate regression model now is of general character, not focused on only *consumer liking* data and *descriptive analysis / sensory profiling* data.

Other than that the concept of computation and presentation of X scores, X & Y correlation loadings, X loadings, Y loadings, explained variance in X and Y is identical to those in the *Prefmap* tab and will not be repeated in this section.

6.7. Conjoint analysis

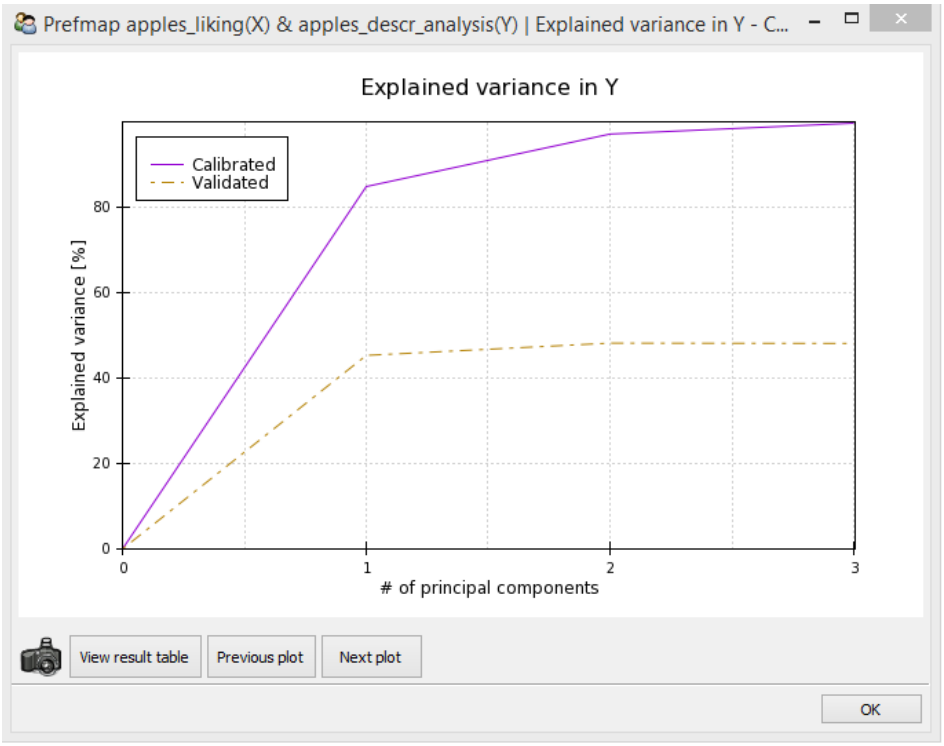


Figure 28: A screenshot of the explained variances in Y *apple descriptive analysis / sensory profiling* data in the preference mapping model.

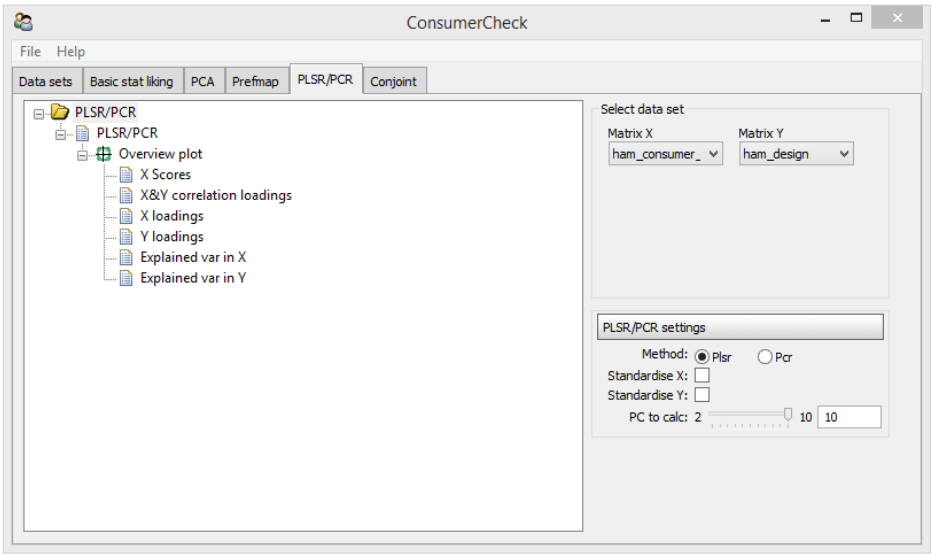


Figure 29: A screenshot of the *PLSR/PCR* tab.

In order to run a conjoint analysis *product design* data, *consumer liking* data and *consumer characteristics* data are required. Fig. 30 shows the screenshot of the conjoint GUI. As usual, model settings may be set with the controls on the right side of the GUI. The *product design* data and *consumer characteristics* data that are used for computation of the conjoint model are selected from their respective drop-down menus. As soon as the data is selected in the drop-down menu its variables are displayed as checkboxes right below. Data of type *consumer liking* appear to the right of the two drop-down menus and may be selected with checkboxes. The reason why a drop-down menu wasn't used for these data is that in some consumer trials one may have asked the consumers to rate their liking on the same set of products for multiple modalities, such as odour, flavour, texture, etc. Having checkboxes rather than one drop-down menu ConsumerCheck allows the user to select multiple *consumer liking* data and to compute multiple conjoint models, i.e. one for each modality. By doing so, this generates multiple tree controls, one for each conjoint model, on the left side of the GUI providing access to multiple conjoint models at once. This may be convenient if the user wants to jump quickly between the models and compare effects of design variables and consumer characteristics for the different liking data. Below the tools for setting the data for conjoint analysis there is another drop-down menu where the user can select the complexity of the conjoint model. A short description of complexity for each conjoint model structure is given below the drop-down menu. In this paper a more detailed description of each conjoint model structure is provided:

Struct 1 The mixed effects model includes fixed main effects. Random effects consist of random consumer effect and interaction between consumer and the main effects.

Struct 2 The mixed effects model includes main effects and all 2-factor interactions. Random effects consist of consumer effect and interaction between consumer and all fixed effects (both main and interaction ones).

Struct 3 This is a full factorial model with all possible fixed and random effects (i.e. including all main effects and all higher-way interactions). The automated reduction in random part is followed by an automated reduction in fixed part. The tests for the random effects use likelihood ratio tests while the tests for the fixed effects use the F-test with Satterthwaite's approximation to degrees of freedom. The automated reduction in the fixed part uses the principle of marginality, i.e. the highest order interactions are tested first: if they are significant, the lower order effects are not eliminated even if being non-significant. This type of structure uses the methodology from Kuznetsova, Christensen, Bavay, and Brockhoff (2015).

There are multiple items at the tree-control from which computational result may be accessed, either in numeric format in tables or as plots. Each of the results provided are described in more detail below. For an illustration the ham data (Fig. 30) were analysed using the conjoint method. The following factors were selected for the model, i.e. they were checked below the drop-down menus: 'product' and 'information' from the *design matrix* as well as 'sex' from the *consumer characteristics*. Age was not included in model since it has too many levels (34), which could cause long computation times or freezing of the software. Finally, *Struct 2* was selected, which means that all selected factors and all their 2-way interactions were included in the model.

Conjoint analysis - LS means

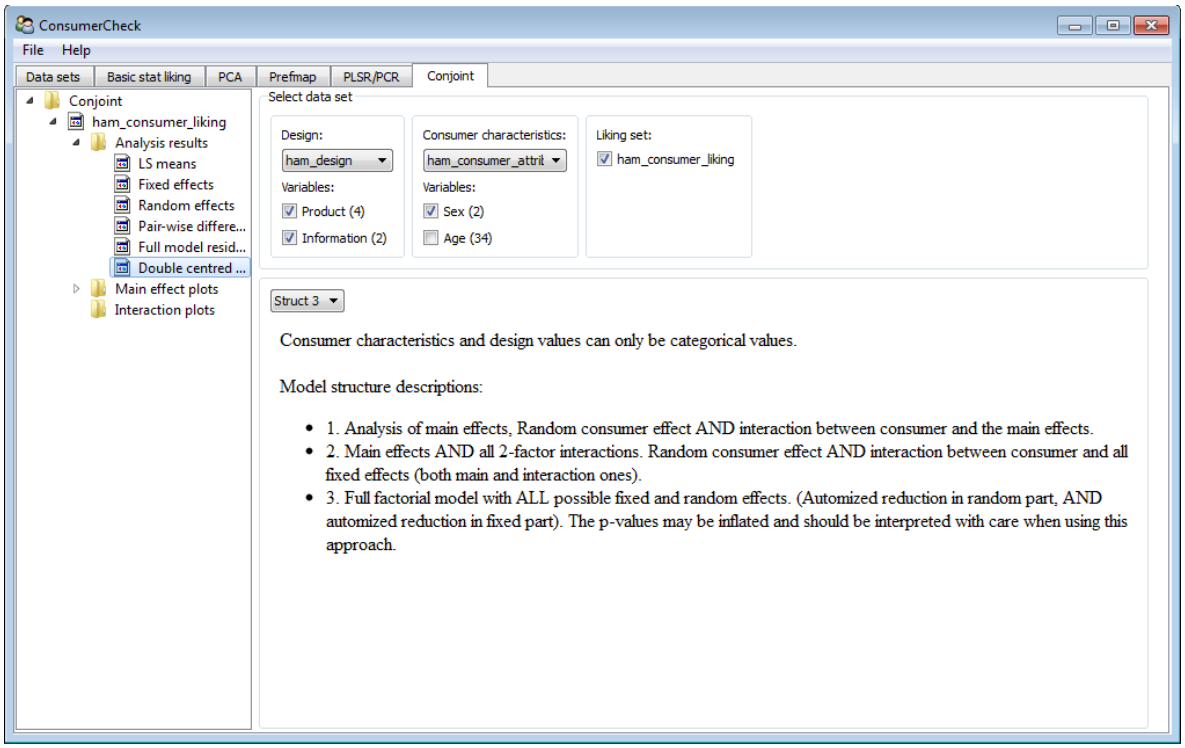


Figure 30: A screenshot of the GUI for conjoint analysis where the *ham product design* data, *ham consumer liking* data and *ham consumer characteristics* data are selected for analysis.

The LS means Table 1 shows population means. In case of balanced data they are exactly the corresponding means. From the column named "Estimate" one may see that the most liked products are Product 3 and 4. Standard errors for the population means, lower and upper 95 percents confidence intervals are also provided.

Conjoint analysis - Fixed effects

Table 2 shows the marginal ANOVA table for fixed effects. Since in this example *Struct 2* was selected, no reduction of the fixed effects was performed. For *Struct 3* the elimination of non-significant effects is performed and an additional column named "elim.num" is provided that shows the order of elimination of effects. From the table it is seen that only main effects for 'Product' and 'Information' seem to be significant: the p value for the 'Product' effect is around 0.01, the p value for the 'Information' effect is slightly higher than 0.05.

Conjoint analysis - Random effects

Table 3 shows an ANOVA-like table for the random effects. Here each random effect was tested with likelihood ratio test. Non-significant random effects were sequentially eliminated if being non-significant according to the default Type 1 error rate 0.1. From the table it is seen that the effect corresponding to interaction between Product and Consumer is highly significant.

Table 1: *LS means* result table.

Model parameter	Information	Product	Sex	Estimate	Standard Error	DF	t-value	Lower CI	Upper CI
Information 1	1	NA	NA	5.6318	0.1398	104.2	40.28	5.3545	5.909
Information 2	2	NA	NA	5.8312	0.1398	104.2	41.71	5.554	6.1085
Product 1	NA	1	NA	5.8084	0.233	309.5	24.93	5.3499	6.2668
Product 2	NA	2	NA	5.1012	0.233	309.5	21.89	4.6428	5.5597
Product 3	NA	3	NA	6.0909	0.233	309.5	26.14	5.6324	6.5493
Product 4	NA	4	NA	5.9256	0.233	309.5	25.43	5.4672	6.384
Sex 1	NA	NA	1	5.8537	0.1831	79	31.97	5.4892	6.2181
Sex 2	NA	NA	2	5.6094	0.1854	79	30.26	5.2404	5.9784
Information:Product 1 1	1	1	NA	5.7287	0.2541	423.3	22.54	5.2292	6.2282
Information:Product 2 1	2	1	NA	5.8881	0.2541	423.3	23.17	5.3885	6.3876
Information:Product 1 2	1	2	NA	4.8981	0.2541	423.3	19.27	4.3986	5.3976
Information:Product 2 2	2	2	NA	5.3043	0.2541	423.3	20.87	4.8048	5.8039
Information:Product 1 3	1	3	NA	5.8754	0.2541	423.3	23.12	5.3759	6.3749
Information:Product 2 3	2	3	NA	6.3063	0.2541	423.3	24.82	5.8068	6.8058
Information:Product 1 4	1	4	NA	6.025	0.2541	423.3	23.71	5.5254	6.5245
Information:Product 2 4	2	4	NA	5.8263	0.2541	423.3	22.93	5.3268	6.3258
Information:Sex 1 1	1	NA	1	5.7073	0.1965	104.2	29.04	5.3176	6.097
Information:Sex 2 1	2	NA	1	6	0.1965	104.2	30.53	5.6103	6.3897
Information:Sex 1 2	1	NA	2	5.5563	0.1989	104.2	27.93	5.1617	5.9508
Information:Sex 2 2	2	NA	2	5.6625	0.1989	104.2	28.46	5.268	6.057
Product:Sex 1 1	NA	1	1	5.8293	0.3275	309.5	17.8	5.185	6.4736
Product:Sex 2 1	NA	2	1	5.4024	0.3275	309.5	16.5	4.7581	6.0468
Product:Sex 3 1	NA	3	1	6.2317	0.3275	309.5	19.03	5.5874	6.876
Product:Sex 4 1	NA	4	1	5.9512	0.3275	309.5	18.17	5.3069	6.5955
Product:Sex 1 2	NA	1	2	5.7875	0.3315	309.5	17.46	5.1352	6.4398
Product:Sex 2 2	NA	2	2	4.8	0.3315	309.5	14.48	4.1477	5.4523
Product:Sex 3 2	NA	3	2	5.95	0.3315	309.5	17.95	5.2977	6.6023
Product:Sex 4 2	NA	4	2	5.9	0.3315	309.5	17.8	5.2477	6.5523

Conjoint analysis - Pairwise differences

Table 4 shows the first part of pairwise comparisons for the fixed factors from Table 2. The last part of the table was omitted because of its length. Column "p-value.adjust" is a p-value with Bonferroni multiple testing correction within each effect. From this table it is seen that e.g. Products 2 and 3 are significantly different from one another.

Conjoint analysis - Full model residuals

These are the residuals of the model that was specified by the user with the *structure* dropdown menu (the model may be a reduced one, depending on the structure the user selected).

Conjoint analysis - Double centred residuals

These are the residuals that are extracted from a mixed effects model with a saturated fixed structure (main effects and all higher-way interactions form fixed part) and one random Consumer effect. These residuals are also called double-centered since they are mean centered for each consumer and they are also mean centered across consumer for each combination of interaction effects. Double-centered residuals can be used for PCA in order to analyse individual differences across consumers according to Endrizzi, Menichelli, Johansen, Olsen, and Naes (2011). This can be done by right-clicking on the *Double centred residuals* branch of the tree control and select *Copy to data sets* from the menu. By doing so the double centred residuals are copied to the *Data sets* tab and are then available for PCA analysis.

Table 2: *Fixed effects* result table.

Model parameters	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
Information	5.24	5.24	1	78.97	3.29	0.073
Product	17.92	5.97	3	236.98	3.82	0.011
Sex	1.38	1.38	1	78.98	0.88	0.351
Information:Product	10.39	3.46	3	236.98	2.2	0.089
Information:Sex	1.13	1.13	1	78.97	0.72	0.399
Product:Sex	1.64	0.55	3	236.98	0.35	0.79

Table 3: *Random effects* result table.

Model parameter	Chi.sq	Chi.DF	p.value
Information:Consumer	1.34	1	0.247
Product:Consumer	167.47	1	<0.001
Consumer	2.2	1	0.138

Conjoint analysis - Main effects plot

Fig. 31 shows population means with their respective 95 percent confidence intervals. From this plot it is seen that Product 3 is the most liked and Product 2 is the least liked. SHOULD WE SAY SOMETHING ABOUT P VALUE AND COLOUR OF FRAME? CAN YOU ADD THIS TORMOD?

Conjoint analysis - Interaction plot

Fig. 32 shows a two-way interaction plot (if interaction effects are part of the fixed structure). From this plot users may also observe whether there is an interaction between factors.

7. Conclusion

ConsumerCheck is an open source data analysis software tailored for analysis of sensory and consumer data. The software comes with a graphical user interface and as such provides non-statisticians and users without programming skills free access to a number of widely used analysis methods within the field of sensory and consumer science. Computational results are presented in plots that are easily generated from the tree-controls within the graphical user interfaces. Since the construction of conjoint analysis models is not always straightforward, ConsumerCheck provides three previously defined model structures of different complexity. ANYTHING ELSE WE SHOULD MENTION HERE?

8. Acknowledgements

We would like to thank the Research Council of Norway and Norwegian food industry for funding the Norwegian part of the ConsumerCheck project. Would also like to thank Direktoratet for Fødevareerhverv and the Danish industry for funding the Danish part of the ConsumerCheck project.

Table 4: Part of the *Pair-wise differences* result table.

Model parameters	Estimate	Standard Error	DF	t-value	Lower CI	Upper CI	p-value	p-value.adjust
Information 1-2	-0.1995	0.1015	319	-1.97	-0.3991	0.0002	0.0502	0.0502
Product 1-2	0.7072	0.3154	237	2.24	0.0858	1.3286	0.0259	0.1554
Product 1-3	-0.2825	0.3154	237	-0.9	-0.9039	0.3389	0.3714	1
Product 1-4	-0.1172	0.3154	237	-0.37	-0.7386	0.5042	0.7105	1
Product 2-3	-0.9896	0.3154	237	-3.14	-1.611	-0.3682	0.0019	0.0114
Product 2-4	-0.8244	0.3154	237	-2.61	-1.4458	-0.203	0.0095	0.057
Product 3-4	0.1652	0.3154	237	0.52	-0.4561	0.7866	0.6009	1
Sex 1-2	0.2443	0.2606	79	0.94	-0.2744	0.7629	0.3514	0.3514
Information:Product 1 1- 2 1	-0.1593	0.203	319	-0.79	-0.5587	0.24	0.433	1
Information:Product 1 1- 1 2	0.8306	0.3465	334.6	2.4	0.149	1.5123	0.0171	0.4788
Information:Product 1 1- 2 2	0.4244	0.3465	334.6	1.22	-0.2573	1.106	0.2216	1
Information:Product 1 1- 1 3	-0.1467	0.3465	334.6	-0.42	-0.8283	0.535	0.6724	1
Information:Product 1 1- 2 3	-0.5776	0.3465	334.6	-1.67	-1.2593	0.1041	0.0965	1
Information:Product 1 1- 1 4	-0.2962	0.3465	334.6	-0.85	-0.9779	0.3854	0.3932	1
Information:Product 1 1- 2 4	-0.0976	0.3465	334.6	-0.28	-0.7792	0.5841	0.7785	1
Information:Product 2 1- 1 2	0.99	0.3465	334.6	2.86	0.3083	1.6716	0.0045	0.126
Information:Product 2 1- 2 2	0.5837	0.3465	334.6	1.68	-0.098	1.2654	0.093	1
Information:Product 2 1- 1 3	0.0127	0.3465	334.6	0.04	-0.669	0.6943	0.9708	1
Information:Product 2 1- 2 3	-0.4183	0.3465	334.6	-1.21	-1.0999	0.2634	0.2283	1
Information:Product 2 1- 1 4	-0.1369	0.3465	334.6	-0.4	-0.8186	0.5448	0.6931	1
Information:Product 2 1- 2 4	0.0618	0.3465	334.6	0.18	-0.6199	0.7435	0.8586	1
Information:Product 1 2- 2 2	-0.4063	0.203	319	-2	-0.8056	-0.0069	0.0462	1
Information:Product 1 2- 1 3	-0.9773	0.3465	334.6	-2.82	-1.659	-0.2956	0.0051	0.1428
Information:Product 1 2- 2 3	-1.4082	0.3465	334.6	-4.06	-2.0899	-0.7266	0.0001	0.0028
Information:Product 1 2- 1 4	-1.1269	0.3465	334.6	-3.25	-1.8085	-0.4452	0.0013	0.0364
Information:Product 1 2- 2 4	-0.9282	0.3465	334.6	-2.68	-1.6098	-0.2465	0.0078	0.2184
Information:Product 2 2- 1 3	-0.571	0.3465	334.6	-1.65	-1.2527	0.1106	0.1003	1
Information:Product 2 2- 2 3	-1.002	0.3465	334.6	-2.89	-1.6836	-0.3203	0.0041	0.1148
Information:Product 2 2- 1 4	-0.7206	0.3465	334.6	-2.08	-1.4023	-0.0389	0.0383	1
Information:Product 2 2- 2 4	-0.5219	0.3465	334.6	-1.51	-1.2036	0.1597	0.133	1
Information:Product 1 3- 2 3	-0.4309	0.203	319	-2.12	-0.8303	-0.0316	0.0345	0.966
Information:Product 1 3- 1 4	-0.1496	0.3465	334.6	-0.43	-0.8312	0.5321	0.6663	1
Information:Product 1 3- 2 4	0.0491	0.3465	334.6	0.14	-0.6326	0.7308	0.8874	1
Information:Product 2 3- 1 4	0.2814	0.3465	334.6	0.81	-0.4003	0.963	0.4174	1
Information:Product 2 3- 2 4	0.4801	0.3465	334.6	1.39	-0.2016	1.1617	0.1669	1
Information:Product 1 4- 2 4	0.1987	0.203	319	0.98	-0.2006	0.598	0.3284	1
Information:Sex 1 1- 2 1	-0.2927	0.1426	319	-2.05	-0.5733	-0.0121	0.041	0.246
Information:Sex 1 1- 1 2	0.1511	0.2796	104.2	0.54	-0.4034	0.7056	0.5902	1
Information:Sex 1 1- 2 2	0.0448	0.2796	104.2	0.16	-0.5097	0.5993	0.873	1
Information:Sex 2 1- 1 2	0.4437	0.2796	104.2	1.59	-0.1108	0.9983	0.1156	0.6936

References

Bates D, Maechler M, Bolker B, Walker S (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7, URL http://CRAN.R-project.org/package=lme4.

Chavent M, Kuentz-Simonet V, Labenne A, Saracco J (2014). "Multivariate analysis of mixed data: The PCAmixdata R package." arXiv:1411.4911v3 [stat.CO] 4 Dec 2014.

Christensen RHB, Brockhoff PB (2014). "sensR—An R-package for sensory discrimination." R package version 1.4-0 http://www.cran.r-project.org/package=sensR/.

Endrizzi I, Menichelli E, Johansen SB, Olsen NV, Naes T (2011). "Handling of individual differences in rating-based conjoint analysis." *Food Quality and Preference*, **22**(3), 241–254. ISSN 09503293, 18736343. doi:10.1016/j.foodqual.2010.10.005.

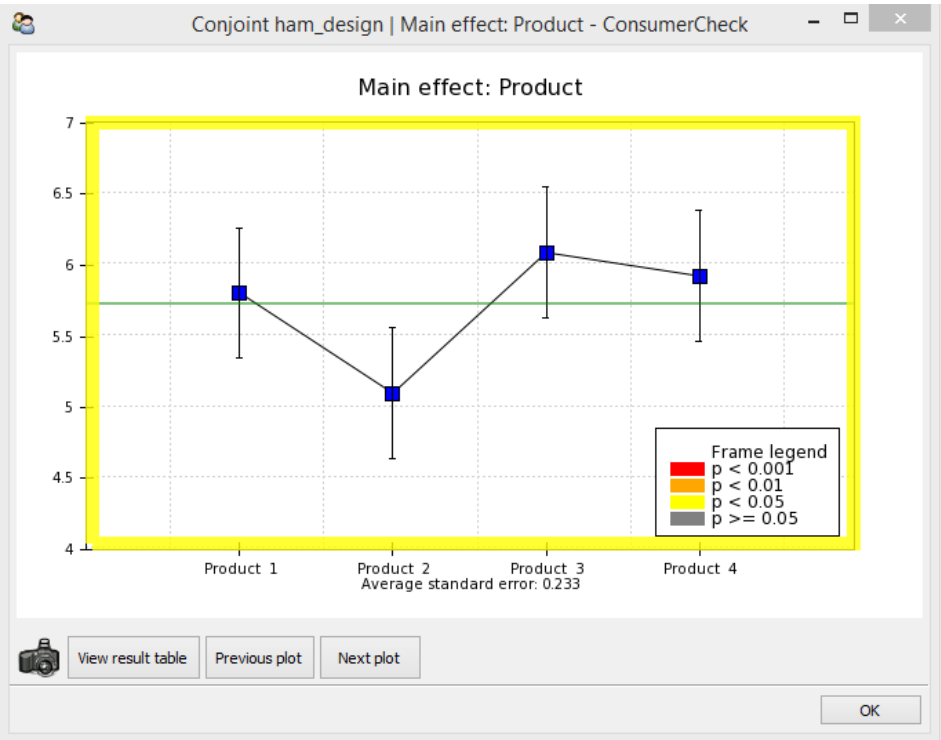


Figure 31: Example of a main effect plot for main effect 'Product' from conjoint analysis of the ham data.

Green P, Rao V (1971). "Conjoint measurement for quantifying judgemental data." *Journal of Marketing Research*, **8**, 355–363.

Green P, Srinivasan V (1978). "Conjoint analysis in consumer research: Issues and outlook." *Journal of Consumer Research*, **5**, 103–123.

Greenhoff K, MacFie H (1994). *Measurements of Food Products*, chapter Preference mapping in practice, pp. 137–166. Glasgow: Blackie Academic and Professional.

Husson F, Le S, Cadoret M (2014). *SensoMineR: Sensory data analysis with R*. R package version 1.20, URL http://CRAN.R-project.org/package=SensoMineR.

Kuznetsova A, Bruun Brockhoff P, Haubo Bojesen Christensen R (2013a). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. R package version 2.0-12, URL http://CRAN.R-project.org/package=lmerTest.

Kuznetsova A, Bruun Brockhoff P, Haubo Bojesen Christensen R (2013b). *SensMixed: Mixed effects modelling for sensory and consumer data*. R package version 2.0-5.

Kuznetsova A, Christensen RH, Bavay C, Brockhoff PB (2015). "Automated mixed ANOVA modeling of sensory and consumer data." *Food Quality and Preference*, **40**, 31–38. ISSN 09503293, 18736343. doi:10.1016/j.foodqual.2014.08.004.

Lawless HT, Heymann H (2010). *Sensory Evaluation of Food - Principles and Practices*. 2nd edition edition. Springer, NY, USA.

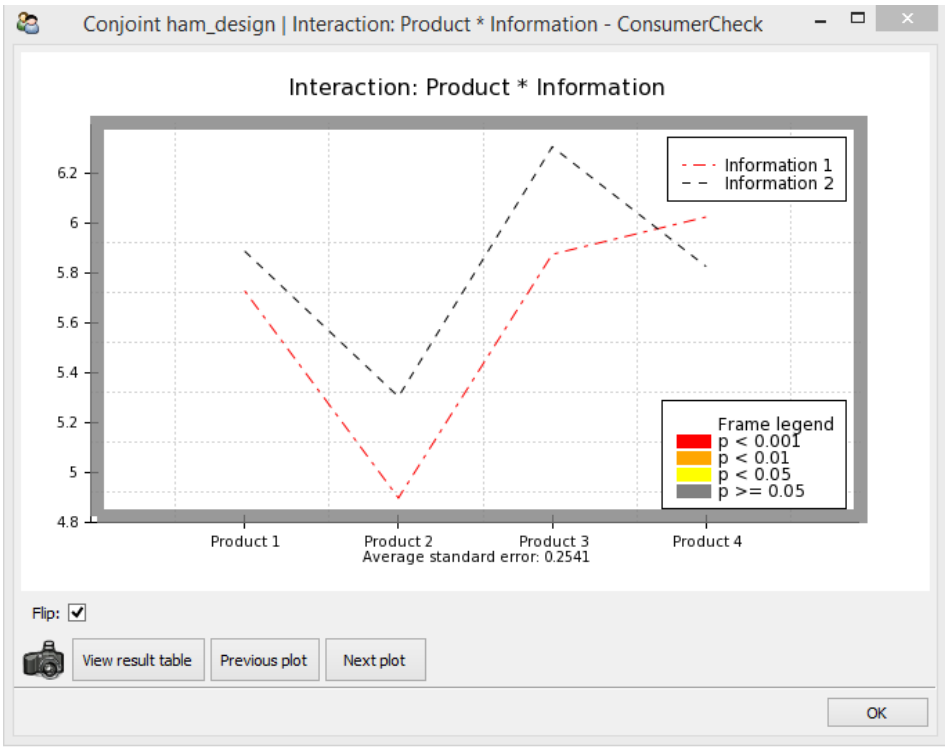


Figure 32: Interaction plot for interaction between main effects 'Information' and 'Product' from conjoint analysis of the ham data.

Mardia K, Kent J, Bibby J (1979). *Multivariate Analysis*. London: Academic Press.

Martens H, Martens M (2001). *Multivariate analysis of Quality: An Introduction*. Wiley, Chichester.

Martens H, Næs T (1988). "Principal components regression in NIR analysis." *Journal of Chemometrics*, **2**, 155–167.

Martens H, Næs T (1989). *Multivariate Calibration*. John Wiley & Sons Ltd, Chichester.

McEwan J (1996). *Multivariate Analysis of Data in Sensory Science*, volume 16 of *Data Handling in Science and Technology*, chapter Preference mapping for product optimization, pp. 71–102. Amsterdam: Elsevier Science B.V.

Næs T, Brockhoff PB, Tomic O (2010). *Statistics for Sensory and Consumer Science*. Wiley, Chichester.

Oliphant TE (2007). "Python for scientific computing." *Computing in Science and Engineering*, **9**, 10–20.

Tomic O, Luciano G, Nilsen A, Hyldig G, Lorensen K, Næs T (2010). "Analysing sensory panel performance in a proficiency test using the PanelCheck software." *European Food Research and Technology*, **230**, 497–211.

Tomic O, Nilsen A, Martens M, Naes T (2007). "Visualization of sensory profiling data for performance monitoring." *LWT–Food Science and Technology*, **40**, 262–269.

Wold H (1982). *Systems under Indirect Observation*, chapter Soft modelling: The basics and some extensions. Amsterdam: North Holland.

Affiliation:

Oliver Tomic
Section for Quality Measurement
Division for Patient Safety and Quality Measurment
Norwegian Knowledge Center for the Health Services
0XXX Oslo, Norway
E-mail: otc@nokc.no
URL: http://eeecon.uibk.ac.at/~zeileis/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd

Appendix E

d-prime interpretation of standard linear mixed model results

Brockhoff P. B., Amorimb I., **Kuznetsova A.**, Søren Bech, Limab R. R.,

delta-tilde interpretation of standard linear mixed model results

Per Bruun Brockhoff[a,*], Isabel de Sousa Amorim[b], Alexandra Kuznetsova[a], Søren Bech[c], Renato Ribeiro de Lima[b]

[a] *DTU Compute, Statistics Section, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark*
[b] *DEX - Departamento de Ciências Exatas, Universidade Federal de Lavras, Campus da UFLA - Caixa Postal 3037 Lavras, MG, Brasil*
[c] *Bang & Olufsen A/S, Struer and Aalborg University, Denmark*

Abstract

We utilize the close link between Cohen's d, the effect size in an ANOVA framework, and the Thurstonian (Signal detection) d-prime to suggest better visualizations and interpretations of standard sensory and consumer data mixed model ANOVA results. The basic and straightforward idea is to interpret effects relative to the residual error and to choose the proper effect size measure. For multi-attribute bar plots of F-statistics this amounts, in balanced settings, to a simple transformation of the bar heights to get them transformed into depicting what can be seen as approximately the average pairwise d-primes between products. For extensions of such multi-attribute bar plots into more complex models, similar transformations are suggested and become more important as the transformation depends on the number of observations within factor levels, and hence makes bar heights better comparable for factors with differences in number of levels. For mixed models, where in general the relevant error terms for the fixed effects are not the pure residual error, it is suggested to base the d-prime-like interpretation on the residual error. The methods are illustrated on a multifactorial sensory profile data set and compared to actual d-prime calculations based on Thurstonian regression modelling through the ordinal package. For more challenging cases we offer a generic "plug-in" implementation of a version of the method as part of the R-package SensMixed. We discuss and clarify the bias mechanisms inherently challenging effect size measure estimates in ANOVA settings.

Keywords:

*Corresponding author. E-mail address: perbb@dtu.dk

Effect Size, Analysis of Variance, F test, *d*-prime

1. Introduction

Data analysis within the sensory and consumer science fields can be particularly challenging due to use of humans as the measurement instrument. Understanding how responses change due to product differences versus change due to subject differences is important. Analysis of variance (ANOVA) is one of the most often employed statistical tools to study differences between products when they are scored by either categorical rating (ordinal) scales and/or unstructured line scales. If for instance one finds that the main product effect is significant, one will be interested in knowing more about which products that are different from each other. To complement the ANOVA F-table, post hoc tests are performed. These procedures, also called multiple comparison tests, are generally based on some correction to protect against having the multiple testing procedure invalidating the overall significance level. Some of the commonly used methods include the Tukey, Bonferroni Newman-Keul's and Ducan's procedures (Næs et al., 2010).

Data analysis based on analysis of variance within the sensory field is usually characterized by a number of such relevant post hoc analyses. To some extend this then handles the effect interpretation part of the analysis. However, it is still valuable to be able to supplement the initial overall ANOVA F-testing, often with highest focus on the p-values with some good measures of overall effect size. In the widely used open source software PanelCheck (Nofima Mat & Ås, 2008) the inbuilt ANOVA results are visualized by multi-attribute bar plots of F-statistics combined with colour coding of the significance results. In this way the F-statistic is used as a kind of effect size measure. This can be a good approach, especially within PanelCheck, where the multi-attribute bar plot of the overall product differences are used only for single-factor product effects and with the same choice of F-test denominator across all the attributes of a plot.

However, the F-statistic itself is generally not the best measure of effect size as it depends on the number of observations for each product. And the various ANOVA mixed models, that we often use for such analysis also complicates the relative effect size handling as generally in mixed models, different effects may have different noise structures, that is, different factors may be tested using different F-test denominators. Moreover, as was pointed

out in Kuznetsova et al. (2015b), it is important, specifically within the sensory and consumer field to be able to also handle more complicated settings than the most simple ones.

More recently, a number of new open source software tools with, among other things, focus on more extended type of mixed model ANOVA for sensory and consumer data have appeared. The ConsumerCheck (Tomic et al., 2015), a tool developed in the same spirit as PanelCheck, offers quite general mixed model analysis of consumer data based on the newly developed more generic R-package lmerTest (Kuznetsova et al., 2014a). In addition, in the still developing R-package SensMixed (Kuznetsova et al., 2015a, 2014b) one of the main purposes is to provide nice and visual multi-attribute interpretations of more complicated analyses. The resulting multi-attribute bar plots will then involve different factors with different number of levels and different number of observations within the levels. It may also involve different mixed model error terms for different factors. All of this calls for some careful thoughts on how to visualize the results of the (mixed) ANOVA results in the best possible way.

The purpose of the present study is to suggest better multi-attribute ANOVA plots for sensory and consumer data based on an effect size expressed in terms of relative pairwise comparisons. We will show how this has a close link to the Thurstonian d-prime, and as such is a generic measure that can be interpreted and compared across any attribute and situation. For balanced data settings, the measure becomes a simple transformation of either one or a few F-statistics making the approach easily applicable for anyone for these cases. For more challenging cases we offer a generic "plug-in" implementation of a version of the method as part of the R-package SensMixed (Kuznetsova et al., 2014b).

The paper is organized such that first, in Section 2, we introduce the basic notion of effect size (ES) in ANOVA framework and the concepts of d-prime. Then in Section 3, we define the effect size $\tilde{\delta}$. Next, in Section 4 it is shown how to estimate the $\tilde{\delta}$ ES measure for certain relevant standard mixed models with possible bias correction. After this, in section 5 we illustrate the method on a multifactorial sensory profile data set and compare the $\tilde{\delta}$ proposed here with the actual d-prime based on Thurstonian modelling. The paper ends with discussions in Section 6.

2. Cohen's *d* and *d*-prime - important effect size measures

Analysis of variance (ANOVA) is one of the most used and the most important methodologies when focus is on investigating product differences in sensory and consumer studies (Næs et al., 2010). ANOVA includes a particular form of null hypothesis statistical testing (NHST) used to identify and to quantify the factors that are responsible for the variability of the response. The null hypothesis for ANOVA is that the means of the factors are the same for all groups. The alternative hypothesis is that at least one mean is different from the others. An F-statistic is obtained in the ANOVA and the F distribution is used to calculate the p-value.

The NHST is a direct form and an easy way to conclude about the statistical significance of a factor, by considering a significance level and a p-value. However, it gets a lot of criticism from researchers of different fields. Yates (1951) observed that researchers paid undue attention to the results of the tests of significance and too little attention to the magnitudes of the effects, which they are estimating. NHST addresses whether observed effects stand out above sampling error by using a test statistic and its p-values, though it is not as useful for estimating the magnitude of these effects (Chow, 1996).

A similar point is made by Sun et al. (2010) and Cohen (1994) phrases it in the following way: "the NHST does not tell us what we want to know, and we so much want to know what we want to know, that, out of desperation, we nevertheless believe that it does!"

The ongoing debate on statistical significance tests has resulted in alternative or supplemental methods for analysing and reporting data. One of the most frequent recommendations is to consider effect size estimates to supplement p-values and to improve research interpretation (Cohen, 1990, 1992, 1994; DeVaney, 2001; Coe, 2002; Steiger, 2004; Cumming & Finch, 2005; Fan, 2010; Sun et al., 2010; Kelley & Preacher, 2012; Grissom & Kim, 2012). Cohen (1990) affirms that the purpose should be to measure the magnitude of an effect rather than simply its statistical significance; thus, reporting and interpreting the effect size is crucial. Fan (2010) shows that p-value and effect size complement each other, but they do not substitute for each other. Therefore, researchers should consider both p-value and effect size.

Cohen (1992) established a relation between the effect size (ES) and NHST definitions: the ES corresponds to the degree in which the H_0 is false, i.e., it is a measure of the discrepancy between H_0 and H_1. Grissom & Kim (2012) states that whereas a test of statistical significance is tradition-

ally used to provide evidence (attained $p-$value) that the null hypothesis is wrong; an ES measures the degree to which such a null hypothesis is wrong (if it is false).

In other words, an effect size is a name given to a family of indices that measure the magnitude of a treatment effect. It can be as simple as a mean, a percentage increase, a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance accounted for. For a two-group setting, the ES quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference (Coe, 2002).

An important class of ES measures is defined by using the standardized effect size. In this class are included the Cohen's d, which is the difference measured in units of some relevant standard deviation (SD) (Cumming & Finch, 2005). Cohen's d is the ES index for the t test of the difference between independent means expressed in units of (i.e., divided by) the within-population standard deviation, which is given by

$$d = \frac{\mu_a - \mu_b}{\sigma}$$

where μ_a and μ_b are independent means and σ is the within-population standard deviation.

There are several effect size measures to use in the context of an F-test for ANOVA. Cohen (1992) defined the effect size for one-way ANOVA as the standard deviation of the K population means divided by the common within-population standard deviation:

$$f = \frac{\sigma_m}{\sigma} \tag{1}$$

where σ_m is the standard deviation of the K population means and σ is the within-population standard deviation.

A very similar measure of standard ES for ANOVA is the root-mean-square standardized effect (Ψ) presented by Steiger (2004). Considering the one-way, fixed-effects ANOVA, in which K means are compared for equality, and there are n observations per group the root-mean-square standardized effect is defined by

$$\Psi = \sqrt{\frac{\frac{1}{K-1}\sum_{i=1}^{K}(\mu_i - \mu)^2}{\sigma^2}} \tag{2}$$

where σ^2 is the mean square error. In fact, this could be just an interpretation of what the Cohen's f really is using $K-1$ for expressing the standard deviation as opposed to using K as others might do. For the remainder of this paper we allow ourselves to consider the Ψ to be our version of the Cohen's standardized ES measure for one-way ANOVA, such that for us "Cohen's $f = \Psi$".

The field of ES measures and estimation thereof is characterized by a certain level of confusion in the choice and use of the various ES measure names, where different names are used for almost the same measures. And, some names are used and defined for population versions of the measures whereas others for sample versions. In addition, the confusion is not diminished by the fact that many of these sample version measures will be biased estimates of the population versions, so often several alternative sample versions of the same population measure exist. It is not the aim of this paper to uncover and review this entire field. Rather we will be clear on exactly how we define the measures we use in both the population versions and the sample versions. We do, however, for completeness, handle, clarify and discuss a number of details related to the various bias relevant mechanisms and investigate what the influence of some of these are. Also, sometimes such ES measures are used for power and sample size computations in the planning phase, and at other times they are used for the actual data analysis. We will use it purely for data analysis interpretation.

Cohen's d for a two-sample setting has a close link with the Thurstonian d-prime, a signal-to-noise ratio from Signal Detection Theory (SDT), which is widely used in sensory science. Mathematically speaking they are exactly the same: the difference between two means relative to a standard deviation. Only the contexts are usually different. The framework of SDT (Green & Swets, 1966) and Thurstonian modelling (Thurstone, 1927) make it possible to investigate the internal and external factors in sensory test and study how these factors influence subjects' test performance (van Hout, 2014).

The SDT approach was first introduced in sensory science with focus on exploring the factors influencing the perceptual process that integrates the information from the senses and the decision process (O'Mahony, 1972, 1979), leading to more effective test designs (O'Mahony, 1995). After this the focus shifted towards understanding and optimizing the decision processes in sensory tests leading to the development of more effective tests that are more predictive of consumer's reality (Hautus et al., 2008).

The Thurstonian approach is responsible for the biggest impact on the

development of sensory difference discrimination test methods (e.g. the duo-trio, triangle, 2-AFC, 3-AFC). Since this kind of tests produce binary data, the statistical methods needed for analysing such data can be found among methods based on the binomial distribution and standard methods for analysing tables of counts. As pointed out by Ennis (1993) one of the weaknesses of working on the count scale is that it is test protocol dependent: the number of expected correct answers for the same products depend heavily on which test protocol that is carried out.

By transforming the number of correct answers into an estimate of the underlying (relative) sensory difference, the Thurstonian model gives the so-called d-prime (d'). The d-prime, which was defined to quantify the effect size, is the estimate of the size of a sensory difference from a particular test. This measure can be seen as generalized measure of sensory difference that expresses size of sensory differences. Since the d' is independent of the test method used, it can be used to accurately and systematically compare sensory tests and study the effects of changes in test design and instructions on the performance of the test (van Hout, 2014). Although maybe the Thurstonian approach is most well-known for its use for sensory discrimination test protocols with binary or ordered categorical outcomes, it has also been suggested and used for ratings data, see e.g. Warnock et al. (2006); Ennis (1999) and also in the context of multivariate analysis of ratings data as e.g. probabilistic multidimensional scaling, cf. (MacKay & Zinnes, 1986). Brockhoff & Christensen (2010) and Christensen et al. (2011) showed how the Thurstonian approach in many cases could be viewed as and embedded into the so-called generalized linear model and/or ordinal regression theory and framework. One benefit of this is the ability to handle regression and ANOVA type analysis within the framework of a Thurstonian approach; where otherwise the most common Thurstonian approach would be to do repeated one- and two-sample computations on various subsets of the data.

3. Methods

We suggest using an ES measure that measures the average pairwise differences between the products or factor levels in question. More specifically, we define it as the root mean square of standardized pairwise differences, which in the balanced one-way ANOVA setting (I groups with n observa-

tions in each group), can be expressed as:

$$\tilde{\delta} = \sqrt{\frac{2}{I(I-1)} \sum_{i_1<i_2}^{I} \left(\frac{\mu_{i_1} - \mu_{i_2}}{\sigma} \right)^2} \tag{3}$$

where $\sum_{i_1<i_2}^{I}$ means the sum of all unique combinations of the two indices. The sum hence includes $I(I-1)/2$ terms, and it is clear that we have expressed the square root of the average of all standardized squared pairwise differences.

The first thing to notice is that the only difference to the Cohen's f or Ψ measures, defined above, is that products - usually in sensory and consumer applications the groups would represent different products - are compared pairwise rather than with the overall mean. This means that we have the following relation between $\tilde{\delta}$ and (our version of) Cohen's f in this balanced one-way ANOVA setting:

$$\tilde{\delta} = \sqrt{2}\Psi = \sqrt{2}f$$

The formal (and short) proof of this is given in the Appendix A.

We need to use our $\tilde{\delta}$ ES measure also for multifactorial settings as this will be an important part of the applications of this. Even though it may be more or less straightforward how this can be done, we believe that it is clarifying to at least express this formally in one of the simplest non-trivial extensions. For the replicated two-factor factorial design, the ANOVA model with main effects of A (α_i, $i = 1, \ldots, I$) and B (β_j, $j = 1, \ldots, J$) and interaction effects A×B (γ_{ij}), we define the $\tilde{\delta}$ ES measures as:

$$\tilde{\delta}_A = \sqrt{\frac{2}{I(I-1)} \sum_{i_1<i_2}^{I} \left(\frac{\alpha_{i_1} - \alpha_{i_2}}{\sigma} \right)^2} \tag{4}$$

$$\tilde{\delta}_B = \sqrt{\frac{2}{J(J-1)} \sum_{j_1<j_2}^{J} \left(\frac{\beta_{j_1} - \beta_{j_2}}{\sigma} \right)^2} \tag{5}$$

$$\tilde{\delta}_{A\times B} = \sqrt{\frac{2}{IJ(IJ-1)} \sum_{ij<i'j'}^{IJ} \left(\frac{\gamma_{ij} - \gamma_{i'j'}}{\sigma} \right)^2} \tag{6}$$

where the sum $\sum_{ij<i'j'}$ means all unique pairwise combinations of all IJ levels. Note, how this definition of the interaction ES measure is a "pure" interaction measure, where indeed all of the many combined levels are compared with each other, but only the real interaction effects are included, that is, the main effects have been removed from this measure. In this way, the size of the interaction ES measure is directly comparable with the size of the main ES measures.

Inspired by the 2-way interaction expression we can formulate a version of $\tilde{\delta}$ that would be applicable for any order of interaction effect $F = F_1 \times F_2 \times \cdots \times F_M$

$$\tilde{\delta}_F = \sqrt{\frac{2}{K(K-1)} \sum_{k<k'} \left(\frac{\gamma_k - \gamma_{k'}}{\sigma} \right)^2} \tag{7}$$

where the sum $\sum_{k<k'}$ means the unique pairwise combination of all combinations of the levels of all the factors in $F = F_1 \times F_2 \times \cdots \times F_M$, and γ_k is the interaction effect for the k'th of all these combinations, where $k = 1, \ldots, K$ and K is the total number of combinations in the interaction effect. In addition, as above: the effects of all lower order effects have then been removed from the measure.

Finally, it is important to realize that these definitions also apply to situations where at the same time we are having yet other effects in the model including the possibility of these being regression (covariate) effects or any combination of such. With this in place we are now ready to begin the discussion of how to compute these measures in practice.

4. The sample estimation of the $\tilde{\delta}$ ES measures

4.1. The independent two- and multigroup one-way ANOVA case

In the two-independent-samples case with $n_1 = n_2 = n$, the absolute value of the pooled t-test statistic is:

$$|t| = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{2}s/\sqrt{n}} = \sqrt{\frac{n}{2}} \frac{|\bar{x}_1 - \bar{x}_2|}{s}$$

where $s = \sqrt{MSE}$ is the pooled standard deviation estimate. When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$. So a simple rescaling of the root-F statistics will correspond to the "plug-in" sample version of $\tilde{\delta}$ in this case (as there is only one term in the sum that defines $\tilde{\delta}$):

$$\sqrt{\frac{2}{n}}\sqrt{F} = \frac{|\bar{x}_1 - \bar{x}_2|}{s}$$

Similarly for the balanced K-group one-way ANOVA setting, we can obtain a "plug-in"-sample estimate of $\tilde{\delta}$ by the same back transformation of the F-statistic:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F}$$

This is almost directly clear from the definition of Ψ above.

4.2. Bias of sample estimates and possible bias corrections

The simple plug-in sample estimate that appeared in a natural way above is in fact not an unbiased estimate of the population $\tilde{\delta}$-value. And even though this may not necessarily prevent us from using such a plug-in approach for the visualization purposes, that are the main focus of the current paper, it is valuable to have some understanding of the bias mechanisms. Some way to possible bias corrections, at least in cases where this will be straightforward, would be valuable.

Assuming the standard normal based one-way ANOVA model, formally the F-statistic

$$F = \frac{MS_{Product}}{MSE}$$

will have a non-central F-distribution as its sampling distribution. The mean of the F can then be found from basic probability and is well-known to be:

$$E(F) = \frac{nK}{nK-2}\left(\frac{n\sum_{i=1}^{K}(\mu_i - \mu)^2/(K-1) + \sigma^2}{\sigma^2}\right) = \frac{nK}{nK-2}\left(\frac{n}{2}\tilde{\delta}^2 + 1\right)$$

where we have re-expressed it in terms of our $\tilde{\delta}$ ES measure. We see that using the plug-in sample estimate by the above given back transformation of the F-statistic will over-estimate the $\tilde{\delta}$ in two ways. Firstly, the fraction $\frac{nK}{nK-2}$ is always larger than 1. This bias mechanism comes from the fact that the

mean of the fraction of two random variables is not the fraction of the means. Since in general for reasonably sized experiments, the number $\sqrt{\frac{nK}{nK-2}}$ will be rather small, this bias will most often not be important, and for most of what we do from here this will be ignored. But if wanted, a simple correction by the factor could be applied.

Secondly, and more importantly, we can see that the less biased back transformation of the F-statistic would be to subtract 1 from it:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F-1}$$

The bias mechanism behind this effect comes from fact that when computing the variability between the sample means we also get some residual error as part of it, which is seen from the classical expected mean square expression for the expected value of the numerator of the F-statistic:

$$E(MS_{product}) = n\sum_{i=1}^{K}(\mu_i - \mu)^2/(K-1) + \sigma^2$$

As this bias can be non-trivial, and the smaller the F, the higher the relative bias, we recommend to correct for this whenever feasible.

4.3. Beyond one-way ANOVA

The back transformation formula we have given above, derived formally in the one-way ANOVA setting, also holds for balanced main effects in multi-factorial settings, but not so for interaction effects. For balanced interaction effects, the proper more general bias-corrected back transformation formula becomes:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{\frac{DF}{K-1}}\sqrt{F-1} \tag{8}$$

where n is the (same) number of observations for each level of the interaction, DF is the degrees of freedom for the interaction effect and K is the number of combined levels of the interaction factor. The proof is given in the Appendix A.

An alternative that can always be used is the "plug-in" method also discussed above: simply extract the relevant effects from the model fit and

then use the defining formula directly. In practice, this can be done e.g. by extracting so-called lsmeans and/or model parameter estimates from the model and use those in the defining formula. This is the approach used in the R-package SensMixed providing a method that works for any setting. The downside of this is the lack of bias correction in the estimates. For purely fixed models we discussed the two main bias mechanisms above. For mixed models a third bias mechanism is discussed below.

4.4. Some standard mixed model sensory and consumer cases

For most sensory and consumer applications the proper model to use would be a mixed model of some kind, where at least effects related to assessors or consumers would be considered random, see e.g. Næs et al. (2010) and Kuznetsova et al. (2015b). Three such examples are the complete consumer preference study corresponding to a completely randomized block setting, the randomized replicated quantitative descriptive sensory analysis (QDA) corresponding to a multi-attribute two-way (products-by-assessor) mixed ANOVA or the batched/sessioned replicated QDA corresponding to a three-way (products-by-assessor-by-batches) mixed ANOVA. These are the three cases that for single factor product study design and complete data can be handled by the PanelCheck tool, leading to either of the following three F-tests for product differences: (using the PanelCheck names for the situations)

$$\begin{aligned}
\text{("No rep")}\ F_{prod} &= \frac{MS_{prod}}{MSE} \\
\text{("2-way")}\ F_{prod} &= \frac{MS_{prod}}{MS_{product\times assessor}} \\
\text{("3-way")}\ F_{prod} &= \frac{MS_{prod}}{MS_{product\times assessor} + MS_{product\times session} - MSE}
\end{aligned}$$

Even though the significance statements in the latter two cases are based on the shown "mixed" error terms, the definition of $\tilde{\delta}$ can be interpreted as measuring the effects relative to the residual standard deviation. Clearly, there could be some potential additional interpretations of considering effect sizes in mixed models relative to other error components than the residual error, but we leave that for future research. In the current paper, we interpret effects relative to the best estimate of the average within individual and within-product variability, that is, the residual error estimate. The "plug-in" method applied to the output of a mixed model, as implemented in

the R-package SensMixed would hence in these simple mixed model settings correspond to the back transformation of the F-statistics from the purely fixed-effect version of the models:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F_{prod,FIXED}} \tag{9}$$

To correct for the bias discussed above, it would be better to use, as above:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F_{prod,FIXED} - 1} \tag{10}$$

However, in the mixed model yet another bias mechanism is operating due to the expected mean squares structure in mixed models. The explicit correction of this bias depends on the actual mixed effect structure. For the two mixed models mentioned above the more proper bias corrections become:

$$\text{("2-way")}\ \hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F_{prod,FIXED} - F_{PA,FIXED}} \tag{11}$$

$$\text{("3-way")}\ \hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F_{prod,FIXED} - F_{PA,FIXED} - F_{PS,FIXED} + 1} \tag{12}$$

where PA standards for product × assessor and PS for product × session. We leave the proofs of why exactly these formulas correct for the mixed model bias to Appendix B.

Such formulas could be straightforwardly deduced for any random component model whenever the expected mean squares structure has been identified. This is straightforwardly done for any sufficiently balanced setting but is more challenging in non-balanced settings. We give here the bias correction formulas for one more situation explicitly, namely the situation corresponding to the example given below: Assume that we have a two-factor product situation (factors P and Q) combined with assuming that only the assessors are random effects, that is, corresponding to the "2-way" error structure above. So all assessors evaluated all product combinations, that is, all combinations of levels of factor P and factor Q. In the full model with all random interaction effects, the bias corrected back transformation formulas are given by:

$$\hat{\tilde{\delta}}^{(P)} = \sqrt{\frac{2}{n}}\sqrt{F_{P,FIXED} - F_{PA,FIXED}} \quad (13)$$

$$\hat{\tilde{\delta}}^{(Q)} = \sqrt{\frac{2}{n}}\sqrt{F_{Q,FIXED} - F_{QA,FIXED}} \quad (14)$$

$$\hat{\tilde{\delta}}^{(PQ)} = \sqrt{\frac{2}{n}}\sqrt{\frac{DF}{K-1}}\sqrt{F_{PQ,FIXED} - F_{PQA,FIXED}} \quad (15)$$

The proof of this is found in Appendix B. Remember, that the fully fixed model should only be run to get the ES measure estimates - everything else, including the significance information, should be extracted from the proper mixed model.

4.5. More general mixed models

For more general mixed models in sensory and consumer applications the product F-statistics can have a more complex form and the effects are estimated by complex weighted averages of the data making the approach of formulating a corresponding fully fixed model followed by a back transformation of a fixed F unfeasible as the effects could be differently estimated in the fixed model. And the bias corrected versions could also be rather challenging to find.

As an alternative approach we suggest instead the general "plug-in" approach. The idea of the "plug-in" method is simply to extract the model parameter estimates of the fixed effects from the software that you are using, in our case most often the lmer function of the lme4 package, (Bates et al., 2014). And then use these to explicitly compute the square root of the average of all possible squared pairwise differences between these. For main effects it amounts to using equations 4 and 5. For 2-way interaction effects equation 6 is used and similarly for higher order interactions.

In the `SensMixed`-package, Kuznetsova et al. (2014b), a version of this "plug-in" approach is implemented, where the so-called LSMEANS, (Harvey, 1975), are used rather than the model parameter estimates. using the LSMEANS implementation in the `lmerTest`-package, (Kuznetsova et al., 2014a). The LSMEANS based $\tilde{\delta}$ estimation implementation can handle main effects, two-way and three-way interactions. Interaction LSMEANS will also include the lower order effects, so for the $\tilde{\delta}$ computation implemented in the `SensMixed`-package these are removed again by subtracting the lower order

effects from the interaction effects using the usual formula, which for a two-way interaction reads:

$$\hat{\gamma}_{ij} = \bar{y}_{ij}^{(LS)} - \bar{y}_{i}^{(LS)} - \bar{y}_{j}^{(LS)} + \bar{y}^{(LS)}$$

The reason for going via the LSMEANS rather than the parameter estimates directly in the `SensMixed`-package implementation is purely for convenience as these were already available. This is not a crucial issue for the current paper, and as any software, it may change in the future.

The "plug-in" method can generally be thought of as corresponding to the simple "F back transformation" method as in some cases, not generally though, it will mathematically exactly be that one, as is clear from the initial discussion of these effect size measures in Section 2. So the price of the "plug-in" is the lack of bias corrections. We still, however, consider this approach much better than the just generalizing the F-plots like used in PanelCheck to more general settings and definitely much better than not doing anything in this respect, as the biases in many cases are expected not to be of major size.

5. Examples

This section will contain an example to illustrate the method on a multifactorial sensory profile data set. We also present a simple example to compare the $\tilde{\delta}$ with the actual d-prime calculations based on Thurstonian regression modelling. The analysis was performed using the SensMixed package. The R-code of the first example is given in the Appendix C. The TVbo data set are available in the SensMixed package.

5.1. Example 1: Multi-way product structures in sensory profile data

The TVbo data set comes from the high-end HIFI company Bang & Olufsen A/S, Struer, Denmark. The main purpose was to test products, specified by two features: Picture (factor with four levels) and TVset (factor with three levels). The 12 combinations of TVset and Picture were assessed by a sensory panel composed by eight trained panelists for a list of 15 different response variables (characteristics of the product) in two replications. The data is available in the SensMixed package named TVbo.

To specify the mixed model, the main effect Assessor plus interactions between Assessor and product effects (TVset and Picture and the interaction

TVset:Picture) are considered random effects. The fixed part contains a multi-way product structure: two main effect TVset and Picture and an interaction between them. The 15 attributes in TVbo data can be analysed all together using the SensMixed package.

In Figure 1, for comparison, a multi-attribute bar plot based on the F-values from the mixed model is presented combined with colour coding of the significance results. Since the mixed model specified here has three fixed effects (TVset, Picture and interaction), the F-tests have different mixed model error terms for each effect. In this way the F-statistic is not comparable because the F-test denominators are different across the attributes. Looking into the multi product structure given by Figure 1 we can see that the main effect TVset is significant for 13 of 15 attributes; the main effect Picture and the interaction are significant for 11 of 15 attributes. For the attributes 2, 4 and 13 for instance, the main effect TVset is significant and Picture is not significant. It means that, for these attributes the products differ mostly due to the effect of TVset. In that way, for the attribute 8 the products differ mostly due to Picture. For the attribute 10, all fixed effects are not significant, that means the assessors were not able to discriminate the products for this attribute. For the remaining attributes, 1, 3, 5, 6, 7, 11, 12, 14 and 15, both main effects are significant and also the interaction, except for the attribute 12. Since the number of levels of the two main effects are different, the F test are not comparable for judging the actual effect sizes.

In Figure 2 the alternative bar plot to visualize the (mixed) ANOVA results is presented based on the $\hat{\tilde{\delta}}$, the effect size measure obtained from the mixed model bias corrected back transformation of the product F-tests coming from the fixed model for TVbo data. Comparing the bars of the delta-tilde plot (Figure 2), it can be seen that the effect of TVset is larger than the effect of Picture for the attributes 1, 2, 4, 5, 6, 7 and 13. The effect of TVset for attribute 6, for instance, is much larger than all the effects for the other attributes. The effect of Picture is larger for the attributes 3, 8, 9, 11, 12, 14 and 15 than for the other attributes. The effect of interaction is larger for the attribute 11 than for the other attributes. It is important to note that the Figure 2 gives us relevant information regarding the size of each effect.

The delta-tilde plot is in several ways a better visual tool, e.g. also when there are F-statistics much larger than the others, e.g. the F-statistic for the TVset effect for the attributes 6 and 15 given in Figure 1. It makes the small

values difficult to visualize. With the back transformation, the ES estimates presented in the Figure 2 has a much smaller range which makes the bar heights better comparable.

[Figure 1 about here.]

[Figure 2 about here.]

Now let us look more closely into an attribute to see how the $\hat{\tilde{\delta}}$ was calculated for each effect. Considering the Attribute 7 as an example. To obtain the bias corrected estimates of $\tilde{\delta}$ we first run the fully fixed effect version of the model and then we apply the back transformation on the product F-test from this model, cf. Table 1. It is important to keep in mind that the fixed effect model is used only to get the product F-tests to apply the back transformation to obtain the ES measure estimates. The significance information should be extracted from the proper mixed model.

[Table 1 about here.]

The back transformation of the F-statistics for the main effect is calculated according to equations (13), (14) and (15):

$$\begin{aligned}
\hat{\tilde{\delta}}_{TV} &= \sqrt{\frac{2}{64}}\sqrt{70.01 - 5.58} = 1.42 \\
\hat{\tilde{\delta}}_{Picture} &= \sqrt{\frac{2}{48}}\sqrt{3.74 - 0.60} = 0.36 \\
\hat{\tilde{\delta}}_{TV*Picture} &= \sqrt{\frac{2}{16}}\sqrt{\frac{6}{11}}\sqrt{4.24 - 1.35} = 0.44
\end{aligned}$$

[Figure 3 about here.]

The delta-tilde estimates for the product effects (TVset, Picture and the interaction TVset:Picture) for the attribute 7 is presented in Figure 3. Since the delta-tilde estimates represents the effect size, the heights of the bars can be comparable between each other. From the Figure 3 we can see that the delta-tilde estimate for TVset is much larger than the others, which means that the effect of TVset is larger than the effect of Picture for the attribute 7. So the impact of TVset effect on the ability to discriminate between products

is higher than the impact of Picture. When interpreting the $\tilde{\delta}$-values we must remember that these are expressing average pairwise differences. This means that if there is only a single product that differs from the rest, say, and the remaining ones are really the same, it will tend to appear as a small average effect in the plot - but potentially still statistically significant. These plots cannot substitute a good post hoc analysis of product differences.

Finally let us compare with what we would have obtained by either of the more biased versions in this case. First the "F-1 back transformation":

$$\begin{aligned}
\hat{\tilde{\delta}}_{TV} &= \sqrt{\frac{2}{64}}\sqrt{70.01-1} = 1.47 \\
\hat{\tilde{\delta}}_{Picture} &= \sqrt{\frac{2}{48}}\sqrt{3.74-1} = 0.34 \\
\hat{\tilde{\delta}}_{TV*Picture} &= \sqrt{\frac{2}{16}}\sqrt{\frac{6}{11}}\sqrt{4.24-1} = 0.47
\end{aligned}$$

and next the "F back transformation" (same as the "plug-in" method in this case):

$$\begin{aligned}
\hat{\tilde{\delta}}_{TV} &= \sqrt{\frac{2}{64}}\sqrt{70.01} = 1.48 \\
\hat{\tilde{\delta}}_{Picture} &= \sqrt{\frac{2}{48}}\sqrt{3.74} = 0.39 \\
\hat{\tilde{\delta}}_{TV*Picture} &= \sqrt{\frac{2}{16}}\sqrt{\frac{6}{11}}\sqrt{4.24} = 0.54
\end{aligned}$$

We see that it does not change the interpretation of the analysis in any important way regarding the comparison of effect sizes between the three effects. The smaller effects the larger the relative biases. And we should remember that also the "mixed model bias corrected" version still is not 100% without bias. The three sets of computations were performed for all the $3 \times 15 = 45$ cases across the 15 attributes. Disregarding the 9 cases with non-significant effects, the gray bars in the plots, the average relative absolute bias for the remaining 36 cases were 8.2% for the "plug-in" alias "F back transformation" method and only 3.3% for the "F-1 back transformation" method. When effects are non-significant, we should not even try to interpret effect sizes. The "F-1 back transformation" method overestimated the effect size for 28 out of the 36 cases and underestimated it for 8 cases.

5.2. *Example 2: Comparison with d-prime from Thurstonian model - simple example*

To compare the $\tilde{\delta}$ with the d-prime from a Thurstonian model we will use the simplest example considering a subset of the TVbo data. The purpose of this is solely to make the explicit point of the close relation between the effects size $\tilde{\delta}$ and the Thurstonian d-prime for readers who might not originally have thought of these two concepts together. Taking the average of TVset1 and TVset2 by Picture for the 8 Assessors we get the subset described in the table 2. Table 3 gives the ANOVA table for the subset of TVbo data.

[Table 2 about here.]

[Table 3 about here.]

The ES measure estimate for this situation is given by the difference between the independent means divided by residual error estimate.

$$\hat{\tilde{\delta}} = \frac{7.05 - 3.8875}{\sqrt{6.27}} = 1.26$$

To calculate the "real" d-prime from Thurstonian model we use the ordinal package (Christensen, 2014). First the subset presented in the table 2 are categorized from 1 to 10, since the response in the cumulative link model (CLM) is usually interpreted as an ordinal response with levels ordered. The categorized data is presented in table 4. Then we obtain the d-prime from the cumulative link model function (see Appendix B) which is equal to 1.26.

[Table 4 about here.]

We can see that the close link between delta-tilde, the effect size in an ANOVA framework, and the Thurstonian d-prime, discussed in the section 2, can be confirmed by a comparison between the real d-prime calculation and the delta-tilde estimate.

6. Summary and Discussion

In this paper we have suggested the use of ES measures as a visual tool to improve the interpretation of the ANOVA table in Analysis of Variance. In spite of having been discussed in literature for decades, ES measures have

not been used extensively for this purpose. Instead more focus has been on the post hoc part of the ANOVA data analysis. We believe that even though the ES plots suggested here cannot substitute a good post hoc analysis, they are valuable additional tools for a good and relevant interpretation of the ANOVA table, and can help to move the focus a bit away from purely looking at p-values but rather focussing on the size of the effects (but still using the p-value information). And this becomes particularly useful in situations with more than a single factor and with several attributes. We have suggested an ES measure that expresses the size of the average pairwise relative differences hence closely related to the d-prime definition in Thurstonian models.

For the mixed model situations that we mostly encounter in sensometrics, we have given three different ways to obtain sample estimates of the measures: The "plug-in" method, the "F-1 back transformation method" and the more elaborate "mixed model bias corrected back transformations". None of these methods are 100% unbiased estimates of the population $\tilde{\delta}$-measure. The "plug-in" method can easily be used based on output from standard mixed model software, and we have implemented this in the R-package SensMixed, and has the strength that it can be used in any complicated setting. For certain simple settings this corresponds to an "F back transformation" and is the more biased version of the three. And especially for small effect sizes it may overestimate the size. The "F-1 back transformation" is easy to use in many simple settings that we often encounter in sensometrics, but is still not fully bias correcting in the mixed models we use. It will most often overestimate the effects size also, but may in some cases actually underestimate.

For the sake of simplicity we still recommend the use of any of these two simple approaches as opposed to simple F-plots or not doing anything in this respect at all. But clearly it would be better to always use the fully bias corrected versions. But we emphasize, that as presented here we suggest to supplement the multi-attribute ANOVAs with a plot. We do not as such suggest to use these measures for the final conclusions of the analysis. The example given here indicate that the level of bias of the biased versions is far from invalidating these approaches, as long as they are only used for the cases with some significance of effects.

Acknowledgments

This work is a part of the Senswell project funded by Innovation Fund Denmark (grant no. 0603-00418B) and the ConsumerCheck project funded by The Danish AgriFish Agency under the Innovation Law (grant no. 3414-08-02347) and the Science without Borders program funded by the Brazilian government agencies CAPES and CNPq. One of the reviewers is highly acknowledged for leading us towards a more comprehensive handling of the mixed model bias issues.

7. Appendix A

Proof of the relation between $\tilde{\delta}$ and Cohen's f in the balanced one-way ANOVA setting:

From the basic relation between the sum of squared deviations from the mean and the sum of squared pairwise differences, that we state here without proof:

$$\sum_{i=1}^{I}(\mu_{-}\bar{\mu})^2 = \sum_{i_1<i_2}^{I}(\mu_{i_1}-\mu_{i_2})^2/I,$$

it follows that the definition of $\tilde{\delta}$:

$$\tilde{\delta} = \sqrt{\frac{2}{I(I-1)}\sum_{i_1<i_2}^{I}\left(\frac{\mu_{i_1}-\mu_{i_2}}{\sigma}\right)^2}$$

also can be expressed as:

$$\tilde{\delta} = \sqrt{\frac{2}{(I-1)}\frac{\sum_{i=1}^{I}(\mu_i-\bar{\mu})^2}{\sigma^2}}$$

Hence we have proved that

$$\tilde{\delta} = \sqrt{2}\Psi = \sqrt{2}f$$

Proof for the more general bias corrected back transformation for balanced interaction effects:

Considering γ_{ij} as the different interaction contributions, the interaction effect is estimated by

$$\hat{\gamma}_{ij} = \bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot}$$

with $K = I \cdot J$ and $n_1 = \cdots = n_K = n$.

$$F \approx \frac{n\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2/DF + \sigma^2}{\sigma^2}$$

Where DF is the interaction degrees of freedom defined by $(I-1)(J-1)$.

$$= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{\gamma_{ij}}{\sigma}\right)^2/DF + 1$$

We use the basic relation between squared paired differences and the sum of square again:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\gamma}_{ij})^2 = \sum_{ij\neq i'j'}^{K}(\hat{\gamma}_{ij} - \hat{\gamma}_{i'j'})^2/(2K)$$

Taking only the lower triangular part of the difference matrix we have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\gamma}_{ij})^2 = \sum_{ij<i'j'}^{K}(\hat{\gamma}_{ij} - \hat{\gamma}_{i'j'})^2/(K)$$

And then

$$F \approx \frac{n}{2}\frac{DF}{K-1}\sum_{ij<i'j'}^{K}\left(\frac{\gamma_{ij}-\gamma_{i'j'}}{\sigma}\right)^2/\left(K(K-1)/2\right) + 1$$

$$= \frac{n}{2}\frac{K-1}{DF}\,(\text{Average squared pairwise dprimes}) + 1$$

And hence, for the interaction we have:

$$\tilde{\delta} = \sqrt{\frac{2}{n}}\sqrt{\frac{DF}{K-1}}\sqrt{F-1}$$

8. Appendix B

"Proofs" of the mixed model bias correction formulas for three specific cases

First we consider what we termed the "2-way" setting: I assessors evaluated J products in R replications with complete data, that is a completely balanced situation in all effects. And we only include product, assessor and the interaction in the model - the first as fixed, the latter two as random effects. In this case standard mixed model theory provides the so-called expected mean squares as: (letting PA denote again the product-by-assessor terms)

$$\begin{aligned} E(MS_{PA}) &= \sigma^2 + R\sigma^2_{PA} \\ E(MS_{prod}) &= \sigma^2 + R\sigma^2_{PA} + IR \cdot \frac{\sigma^2 \cdot \tilde{\delta}^2}{2} \end{aligned}$$

where we for the fixed product effect term has expressed exactly how this is related to our effect size measure $\tilde{\delta}$, which comes from Appendix A. Now it becomes immediately apparent that subtracting the interaction F statistic (rather than just "1") is the better bias correction:

$$E(F_{prod,FIXED} - F_{PA,FIXED}) \approx \frac{1}{\sigma^2}\left(E(MS_{prod}) - E(MS_{PA})\right) = IR \cdot \frac{\tilde{\delta}^2}{2}$$

And we have "proved" that equation 11 provides a proper bias corrected version of the given sample estimate of $\tilde{\delta}$. We emphasize again that due to the first bias mechanism discussed in the paper (the nonlinearity of the fraction function), this is still not a 100% unbiased estimate. And in fact the square root transformation itself also induces yet another, so far undiscussed, bias into the system as this is also not a linear function. However, this kind of unbiasedness is a quite commonly accepted one, as the basic sample standard deviation suffers from the same unbiasedness in estimating the population standard deviation. We mention this here to stress that one should not necessarily get too worried about the fact that certain statistics are biased, as long as the bias in practical situations is not important.

The second case, the "3-way" setting where again I assessors evaluated J products in R replications with complete data, that is a completely balanced situation in all effects. But now we also include the "Replication" factor together with the two additional interactions, all as random effects in the

model. The relevant expected mean squares are now:

$$\begin{aligned} E(MS_{PA}) &= \sigma^2 + R\sigma^2_{PA} \\ E(MS_{PR}) &= \sigma^2 + I\sigma^2_{PR} \\ E(MS_{prod}) &= \sigma^2 + R\sigma^2_{PA} + I\sigma^2_{PR} + IR \cdot \frac{\sigma^2 \cdot \tilde{\delta}^2}{2} \end{aligned}$$

Now we can immediately see, without repeating all the details from above, that subtracting both $F_{PA,FIXED}$ and $F_{PR,FIXED}$ will subtract the the two relevant terms but also subtract the residual error σ^2 twice and we have to add the 1 to get that one in again. And we have similarly "proved" equation 12.

Finally, the case with a 2-factor product structure (factors P and Q) combined with just incorporating the assessor random effect is considered: So I assessors evaluated all $J_P \times J_Q$ product combinations in R replications with complete data. In the mixed model we include P, Q and PQ as fixed effects and A, PA, QA and PQA as random effects. The relevant expected mean squares are now:

$$\begin{aligned} E(MS_{PQA}) &= \sigma^2 + R\sigma^2_{PQA} \\ E(MS_{PA}) &= \sigma^2 + R\sigma^2_{PQA} + RJ_Q\sigma^2_{PA} \\ E(MS_{QA}) &= \sigma^2 + R\sigma^2_{PQA} + RJ_P\sigma^2_{QA} \\ E(MS_{P}) &= \sigma^2 + R\sigma^2_{PQA} + RJ_Q\sigma^2_{PA} + Q(P) \\ E(MS_{Q}) &= \sigma^2 + R\sigma^2_{PQA} + RJ_P\sigma^2_{QA} + Q(Q) \\ E(MS_{PQ}) &= \sigma^2 + R\sigma^2_{PQA} + Q(PQ) \end{aligned}$$

where we have adopted the classical notation "Q(A)" for the fixed effect contribution of fixed factor A in the expressions, since the details of these terms follow from above. And now the correction formulas 13, 14 and 15 follow directly as above: the subtracted F-statistics will in all three cases remove exactly the bias mixed model terms.

9. Appendix C

R-code for the analysis of the TVbo data set in SensMixed package

The mixed model for one attribute y_{ijkl} can be specified as:

$$\begin{aligned} y_{ijkl} = \mu + \tau_j + \rho_k + \gamma_{jk} \\ + a_i + b_{ij} + c_{ik} + d_{ijk} + \epsilon_{ijkl} \end{aligned} \tag{16}$$

$$\begin{aligned} a_i &\sim N(0, \sigma^2_{assessor}) \\ b_{ij} &\sim N(0, \sigma^2_{assessor \times TVset}) \\ c_{ik} &\sim N(0, \sigma^2_{assessor \times Picture}) \\ d_{ijk} &\sim N(0, \sigma^2_{assessor \times TVset \times Picture}) \\ \epsilon_{ijkl} &\sim N(0, \sigma^2_{error}) \end{aligned}$$

The fixed part of the model contains a multi-way product structure given by τ_j, ρ_k and γ_{jk} that represents the effect of TVset and Picture and the interaction TVset:Picture respectively. The random part of the model is consisting of the main effect Assessor represented by a_i plus the interactions between Assessor and the fixed effects (TVset and Picture and the interaction TVset:Picture) given by b_{ij}, c_{ik} and d_{ijk}.

Using the SensMixed package we can analyse the 15 attributes in TVbo data with a few command lines. First we attach the SensMixed package (version 2.0-7) by typing the following command in the R console:

```
library(SensMixed)
```

The TVbo data set is available in the SensMixed package. To access the TVbo data use the command:

```
data(TVbo)
```

Then we use the function sensmixed to construct the mixed model for all attributes:

```
resTV <- sensmixed(attributes=names(TVbo)[5:ncol(TVbo)],
                        Prod_effects=c("TVset", "Picture"),
                        individual="Assessor",
                        calc_post_hoc = TRUE,
                        product_structure=3,
```

```
                        error_structure ="No_Rep",
                        reduce.random=FALSE,
                        parallel=FALSE,
                        data=TVbo)
```

The sensmixed function contains a lot of arguments, here we explain the arguments used above:

- attributes: a vector containing the names of the sensory attributes
- Prod_effects: names of the variables related to the product
- individual: name of the column in the data that represent assessors
- data: data frame (data from sensory studies)
- product_structure: one of the values in $1, 2, 3$.
  - 1: only main effects will enter the initial model.
  - 2: main effects and 2-way interaction.
  - 3: all main effects and all possible interaction.
- error_structure ="No_Rep": assessor effect and all possible interactions between assessor and product effects.

The mixed models for each attribute are constructed using the lme4 package (Bates et al., 2014) and then the step method from the lmerTest (Kuznetsova et al., 2014a) is applied to each model. By default the non-significant random effects are eliminated from the model according to the specified by a user Type 1 error (Kuznetsova et al., 2015b). However to estimate the delta-tilde and compare the bars of the plot, the elimination of the random effects is not required. It can be done by the argument reduce.random=FALSE. By default the computation is done in parallel Kuznetsova et al. (2015b). Here we chose parallel=FALSE.

The sensmixed function provides us with the tables of the random and fixed part of the model as well the bar plot presented in the section 5. To get the results we simply type the following into R console:

```
resTV
plot(resTV, dprime=TRUE, isRand = FALSE)
```

This will then produce the "plug-in" version, that is, the "F back transformation" version and hence will be slightly, but not visually different from the bias corrected one given in the paper in any important way.

R-code to obtain the *d*-prime from Ordinal package

First we attach the packages by the typing the following command in the R console:

```
library(ordinal)
```

Then categorize the subset of TVbo data:

```
TVbo$Cutting_ord1=
    as.integer((cut2((TVbo$A7-1)/(max(TVbo$A7)-1),
    cuts=(0:10)/10)))
TVbo$Cutting_ord2=factor(TVbo$Cutting_ord1,ordered=TRUE)
```

And finally use the clm function to obtain the *d*-prime estimate:

```
clm <- clm(Cutting_ord2 ~ TVset,link="probit",
            data = subset(TVbo, TVset != "TV2"))
coef(clm)[-(1:9)]
round(coef(clm),2)
```

References

Amorim, I. S., Kuznetsova, A., de Lima, R. R., Christensen, R. H. B., & Brockhoff, P. B. (2014). Sensmixed - an r package for mixed effects modelling for sensory and consumer data. In *EuroSense 2014: A Sense of Life*. Copenhagen: Elsevier.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. URL: http://CRAN.R-project.org/package=lme4 r package version 1.1-7.

Brockhoff, P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, *21*, 330338. doi:10.1016/j.foodqual.2009.04.003.

Chow, S. (1996). *Statistical Significance: Rationale, Validity and Utility*. Introducing Statistical Methods Series. SAGE Publications. URL: http://books.google.com.br/books?id=0DWeqYsehDsC.

Christensen, R. H. B. (2014). *ordinal - Regression Models for Ordinal Data*. R package version 2014.11-14 http://www.cran.r-project.org/package=ordinal/.

Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2011). Statistical and thurstonian models for the a-not a protocol with and without sureness. *Food Quality and Preference*, *22*, 542549. doi:10.1016/j.foodqual.2011.03.003.

Coe, R. (2002). It's the effect size, stupid. what effect size is and why it is important. *Annual Conference of the British Educational Research Association*, *1*, 12–14. URL: http://www.leeds.ac.uk/educol/documents/00002182.htm.

Cohen, J. (1990). Things i have learned (so far). *American Psychologist*, *45(12)*, 1304–1312. doi:10.1037/0003-066X.45.12.1304.

Cohen, J. (1992). A power prime. *Psychological Bulletin*, *112(1)*, 155–159. doi:10.1037/0033-2909.112.1.155.

Cohen, J. (1994). The earth is round (p ¡ .05). *American Psychologist*, *49(12)*, 997–1003. doi:10.1037/0003-066X.49.12.997.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60(2)*, 170–180. doi:10.1037/0003-066X.60.2.170.

DeVaney, T. A. (2001). Statistical significance, effect size, and replication: What do the journals say? *The Journal of Experimental Education*, *69*, 310–320. URL: http://www.jstor.org/stable/20179992.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, *8*, 353–370. doi:10.1111/j.1745-459X.1993.tb00225.x.

Ennis, D. M. (1999). Thurstonian models for intensity ratings. *IFPress*, *2 (3)*, 2 – 3. URL: http://ifpress.com/publications-cat/technical-reports/ifpress-23-thurstonian-models-for-intensity-ratings/.

Fan, X. (2010). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94-5*, 275–282.

Green, D. M., & Swets, J. A. (1966). *Effect sizes for research: univariate and multivariate applications*. (1st ed.). John Wiley & Sons Ltd.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications*. (2nd ed.). Taylor & Francis Group, LLC.

Harvey, W. R. (1975). Least-squares analysis of data with unequal subclass numbers. *Agricultural Research Service*, *H4*.

Hautus, M. J., O'Mahony, M., & Lee, H. S. (2008). Decision strategies determined from the shape of the same-different roc curve: what are the effects of incorrect assumptions? *Journal of Sensory Studies*, *23*, 743–764. doi:10.1111/j.1745-459X.2008.00185.x.

Kelley, K., & Preacher, K. J. (2012). On effect size. *American Psychological Association*, *17 (2)*, 137 – 152. doi:10.1037/a0028086.

Kuznetsova, A., Amorim, I. S., Christensen, R. H. B., Lima, R. R., & Brockhoff, P. B. (2015a). Analysing sensory data in a mixed effects model framework using the r package sensmixed. *to be submitted to Food Quality and Preference*, .

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014a). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).*. URL: http://R-Forge.R-project.org/projects/lmertest/ r package version 2.0-11/r63.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014b). *SensMixed: Mixed effects modelling for sensory and consumer data*. URL: http://R-Forge.R-project.org/projects/lmertest/ r package version 2.0-5/r68.

Kuznetsova, A., Christensen, R. H. B., Bavay, C., & Brockhoff, P. B. (2015b). Automated mixed anova modeling of sensory and consumer data. *Food Quality and Preference*, *40*, 31–38. doi:10.1016/j.foodqual.2014.08.004.

MacKay, D. B., & Zinnes, J. L. (1986). A probabilistic model for the multidimensional scaling of proximity and preference data. *Marketing Science*, *5 (4)*, 325–344. doi:10.1287/mksc.5.4.325.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for Sensory and Consumer Science*. Chichester, UK: John Wiley and Sons Ltd. doi:10.1002/9780470669181.fmatter.

Nofima Mat, N., & Ås (2008). Panelcheck software. ¡www.panelcheck.com¿.

O'Mahony, M. (1972). Salt taste sensitivity: a signal detection approach. *Perception*, *1(4)*, 459 – 464. doi:10.1068/p010459.

O'Mahony, M. (1979). Short-cut signal detection measures for sensory analysis. *Journal of Food Science*, *44*, 302–303. doi:10.1111/j.1365-2621.1979.tb10071.x.

O'Mahony, M. (1995). Who told you the triangle test was simple? *Food Quality and Preference*, *6*, 227 – 238. doi:10.1016/0950-3293(95)00022-4.

O'Mahony, M., & Hautus, M. J. (2008). The signal detection theory roc curve: Some applications in food sensory science. *Journal of Sensory Study*, *23*, 186 – 204. doi:10.1111/j.1745-459X.2007.00149.x.

Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9(2)*, 164–182. doi:10.1037/1082-989X.9.2.164.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102(4)*, 989–1004. doi:10.1037/a0019507.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34(4)*, 273–286. doi:10.1037/h0070288.

Tomic, O., Brockhoff, P. B., Kuznetsova, A., & Næs, T. (2015). Consumercheck: a software for analysis of sensory and consumer data. *Journal of Statistical Software*, *in press*.

van Hout, D. (2014). *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*. Phd thesis University Rotterdam, Erasmus Research Institute of Management (ERIM). ISBN 978-90-5892-350-9.

Warnock, A. R., Shumaker, A. N., & Delwiche, J. F. (2006). Consideration of thurstonian scaling of ratings data. *Food Quality and Preference*, *17*, 556 – 561. doi:10.1016/j.foodqual.2006.03.003.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46, Issue 253*, 19–34. doi:10.1080/01621459.1951.10500764.

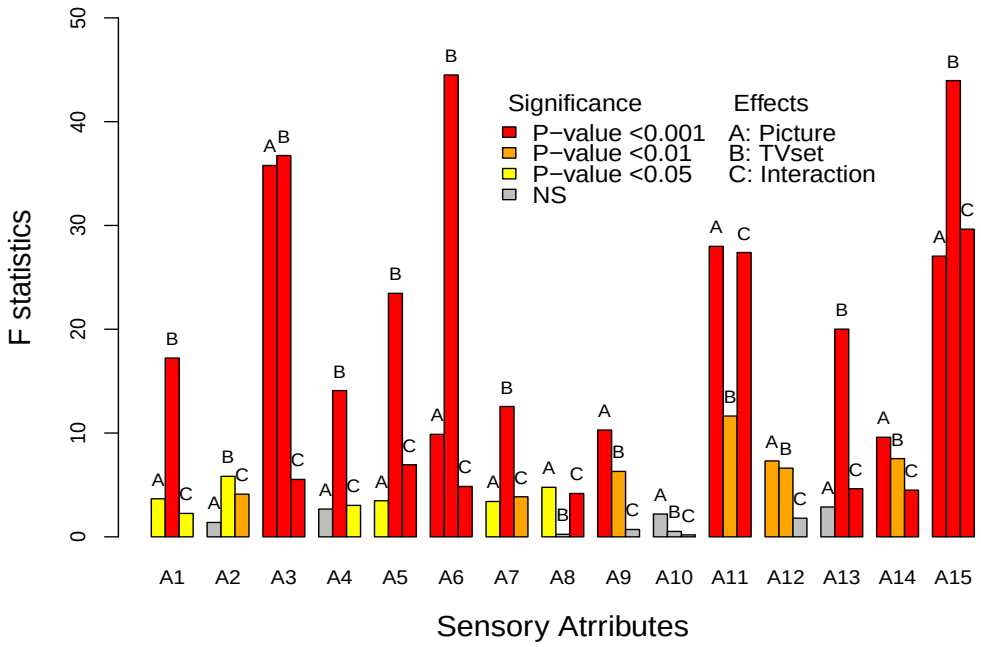


Figure 1: Bar plot for F values for fixed effects of TVbo data.

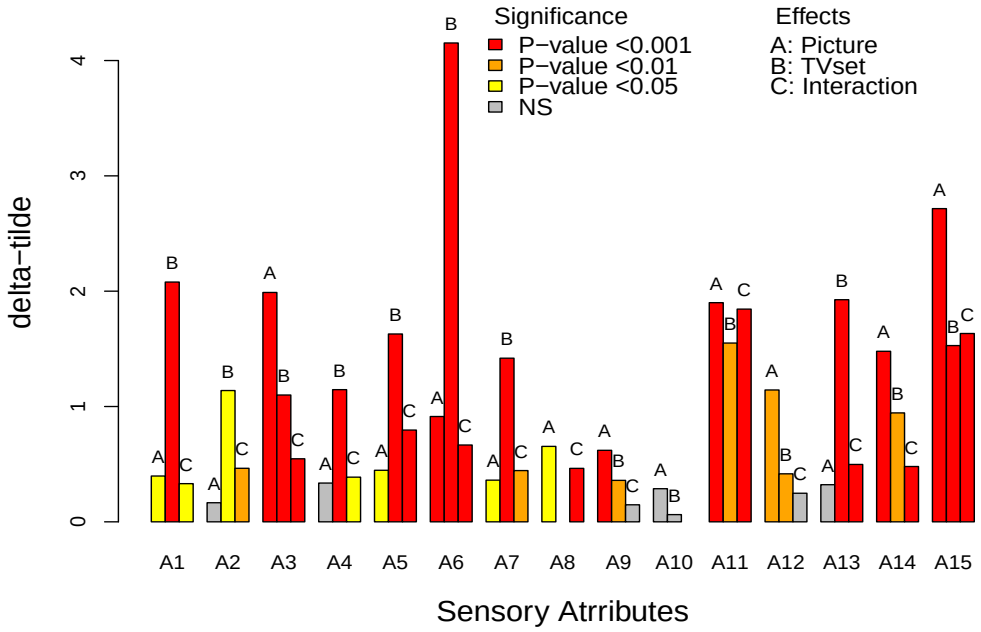


Figure 2: Bar plot based on delta-tilde for fixed effects of TVbo data - the mixed model bias corrected version.

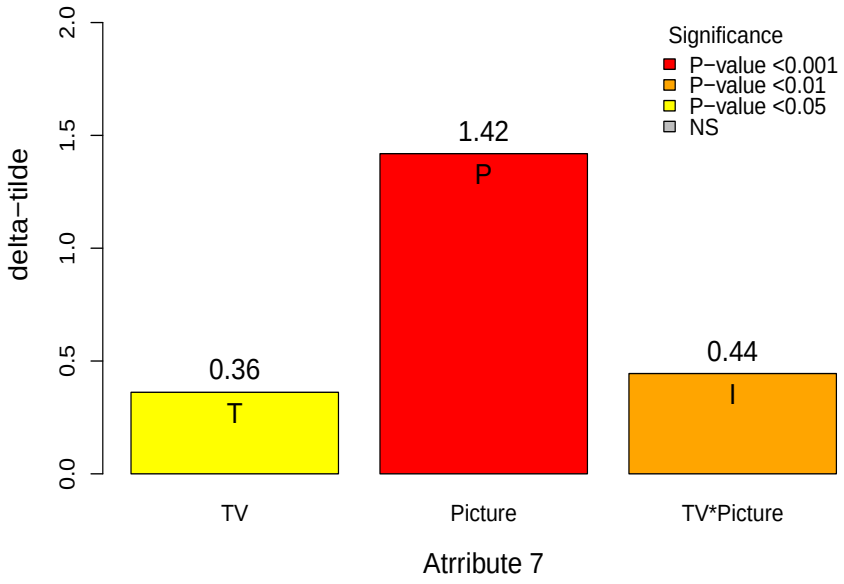


Figure 3: Bar plot for delta-tilde based on F-statistics from fixed effects model for attribute 7.

Table 1: ANOVA table for the fixed effect model for attribute 7 of TVbo data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TVset	2	247.41	123.70	70.01	0.0000
Picture	3	19.84	6.61	3.74	0.0136
TVset:Picture	6	44.98	7.50	4.24	0.0007
Assessor	7	130.87	18.70	10.58	0.0000
TVset:Assessor	14	137.93	9.85	5.58	0.0000
Picture:Assessor	21	22.51	1.07	0.60	0.9047
TVset:Picture:Assessor	42	99.83	2.38	1.35	0.1183
Residuals	96	169.64	1.77		

Table 2: Subset of TVbo data

Assessor	1	2	3	4	5	6	7	8	Mean
TVset1	3.7	5.5	7.1	2.6	11.3	9.1	8.8	8.3	7.0500
TVset2	1.4	4.0	4.8	2.1	7.4	5.2	2.1	4.1	3.8875

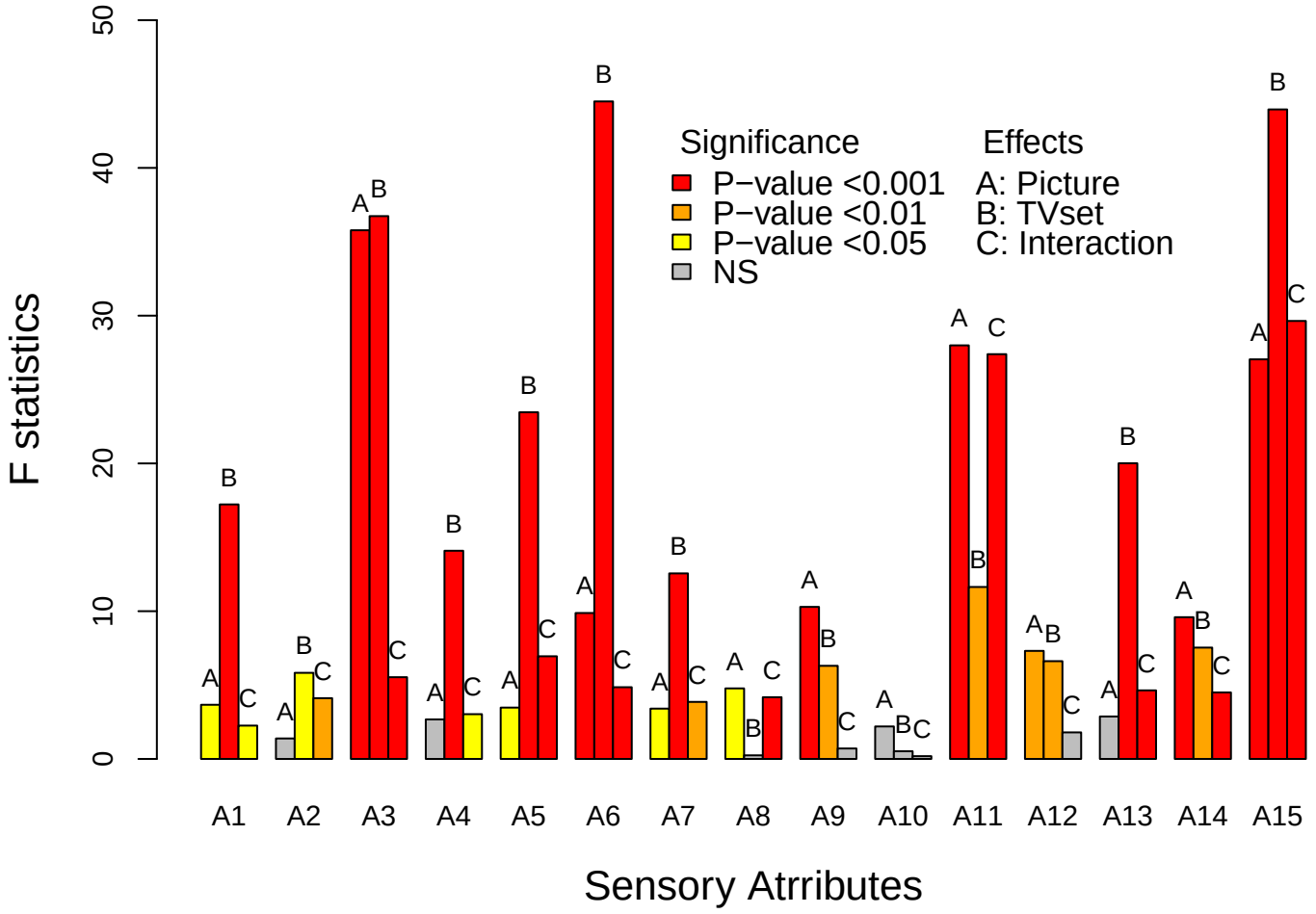
Table 3: ANOVA table for subset of TVbo data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tvset	1	40.01	40.01	6.38	0.0243
Residuals	14	87.85	6.27		

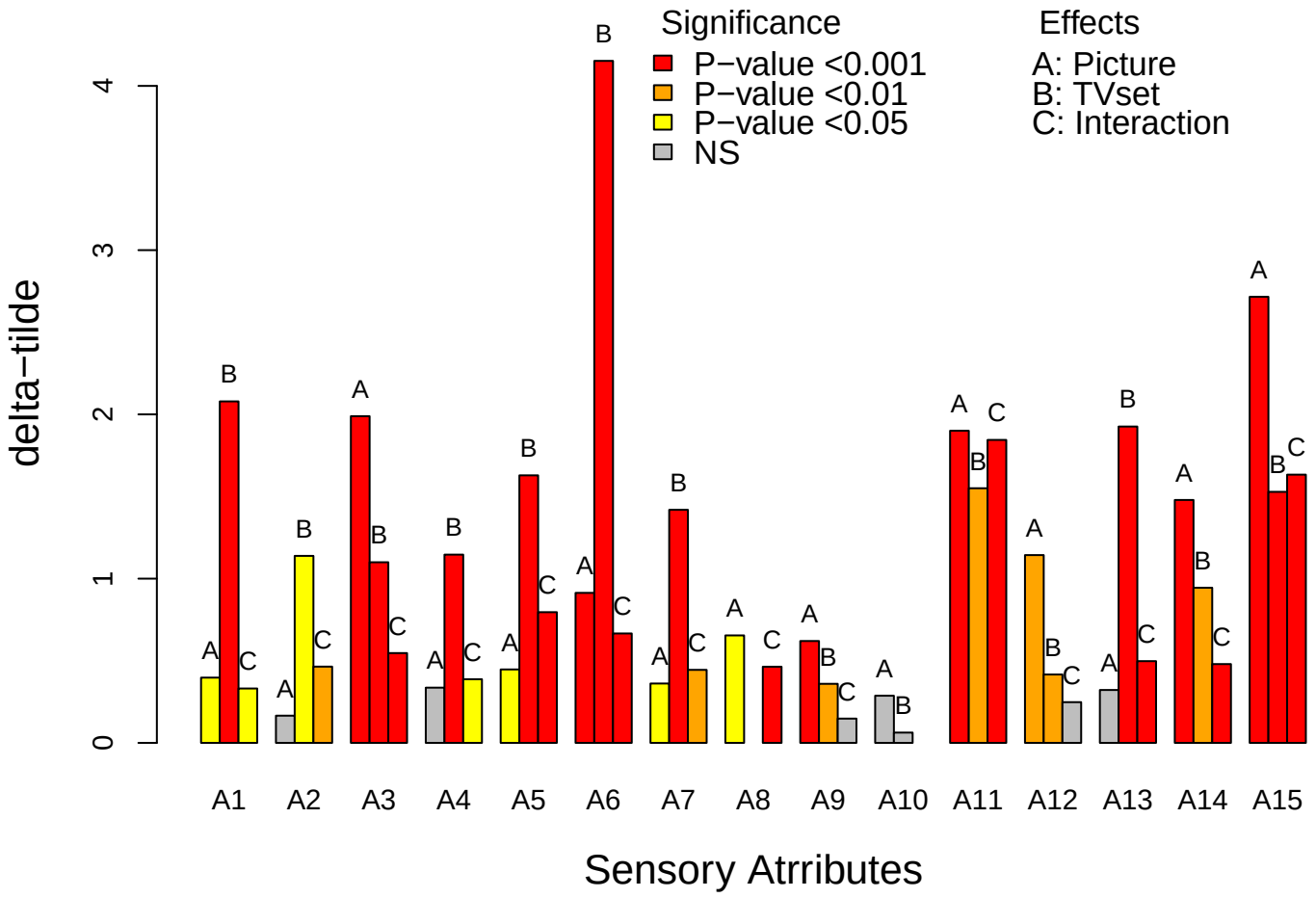
Table 4: Categorized data for subset of TVbo data

Assessor	1	2	3	4	5	6	7	8
TVset1	3	5	6	2	10	8	8	8
TVset2	1	3	4	2	7	5	2	4

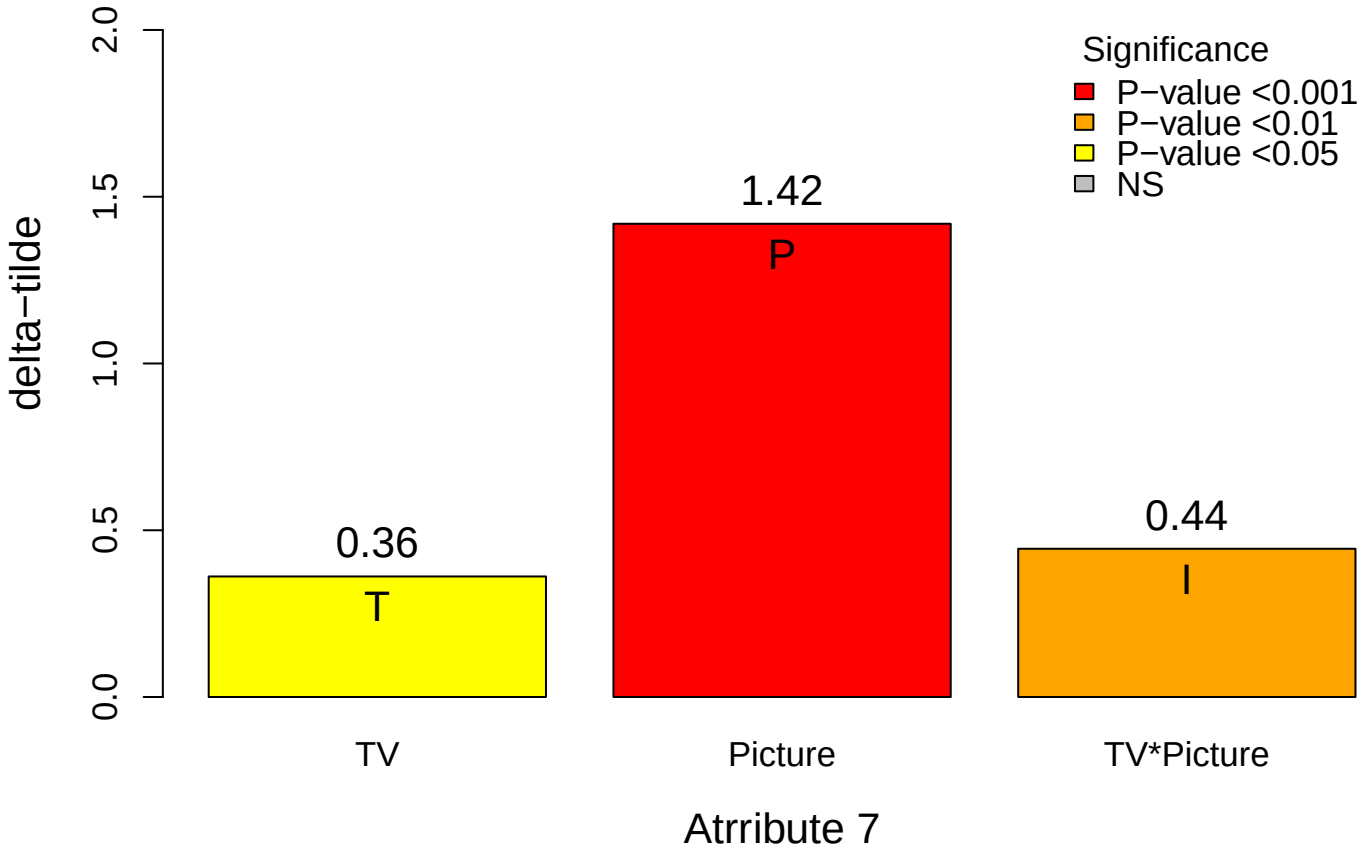
Figure


F statistics
0
10
20
30
40
50
Significance
P−value <0.001
P−value <0.01
P−value <0.05
NS
Effects
A: Picture
B: TVset
C: Interaction
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
Sensory Atrributes

Figure


Significance
P−value <0.001
P−value <0.01
P−value <0.05
NS
Effects
A: Picture
B: TVset
C: Interaction
delta−tilde
0
1
2
3
4
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
Sensory Atrributes

Figure


delta−tilde
0.0
0.5
1.0
1.5
2.0
0.36
T
1.42
P
0.44
I
TV
Picture
TV*Picture
Atrribute 7
Significance
P−value <0.001
P−value <0.01
P−value <0.05
NS

Appendix F

Consideration of Sample Heterogeneity and in-Depth Analysis of Individual Differences in Sensory Analysis

Bavay, Cécile, Per B. Brockhoff, Alexandra Kuznetsova, Isabelle Maître, Emira Mehinagic, and Ronan Symoneaux. 2014. "Consideration of Sample Heterogeneity and in-Depth Analysis of Individual Differences in Sensory Analysis." Food Quality and Preference 32: 126–31. doi:10.1016/j.foodqual.2013.06.003.

Food Quality and Preference 32 (2014) 126–131

Contents lists available at SciVerse ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Consideration of sample heterogeneity and in-depth analysis of individual differences in sensory analysis

Cécile Bavay [a,*], Per Bruun Brockhoff [b], Alexandra Kuznetsova [b], Isabelle Maître [a], Emira Mehinagic [a], Ronan Symoneaux [a]

[a] LUNAM Université, SFR QUASAV 4207, Groupe ESA, UPSP GRAPPE, 55 Rue Rabelais BP 30748, F-49007 Angers Cedex 01, France
[b] DTU Informatics, Statistical Section, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kongens Lyngby, Denmark

ARTICLE INFO

Article history:
Received 3 October 2012
Received in revised form 20 February 2013
Accepted 7 June 2013
Available online 14 June 2013

Keywords:
Assessor model
Mixed model
Sample heterogeneity
Sensory profile
Variability

ABSTRACT

In descriptive sensory analysis, large variations may be observed between scores. Individual differences between assessors have been identified as one cause for these variations. Much work has been done on modeling these differences and accounting for them through analysis of variance (ANOVA). When the products studied are prone to biological heterogeneity (e.g. fruits, vegetables, cheeses, etc.), variations in the data may be due to assessor differences and/or product heterogeneity. The present paper proposes an approach for quantifying these two sources of variation. For individual differences, an extended version of the assessor model approach is applied. The data set used in the paper is based on sensory evaluations of three apple samples scored by a panel of 19 assessors using seven descriptors in four replicates. The application of the extended assessor model approach to unbalanced data provides more insight into assessor differences and a better test for product differences. These results demonstrate the importance of choosing the right model and taking all potential sources of variation into account.

1. Introduction

Sensory quality is commonly assessed through conventional sensory profiling methods. The data resulting from such methods present variations that may be due to assessor differences and/or sample heterogeneity. On one hand, individual differences between assessors are an inherent source of variation. For example, assessors may vary in both their perception and their use of the intensity scale. They may differ in their average (level), in their dispersion on the scale (scaling), in their repeatability (variability) and even in their ranking of the products (disagreement) (Brockhoff, 2003). Training may reduce, but not erase, these effects. This issue of assessor differences has been investigated by many authors (Brockhoff, 1998, 2003; Brockhoff & Skovgaard, 1994; Næs, 1990, 1998; Næs & Solheim, 1991; Romano, Brockhoff, Hersleth, Tomic, & Næs, 2008; Schlich, 1994, 1996). On the other hand, products such as fruits, vegetables, cheeses, etc. are prone to biological heterogeneity. Many authors have highlighted this issue in studies about apples. For example, in a study to develop a specific sensory methodology for the assessment of Cox's Orange Pippin apples, Williams and Carter (1977) reported difficulties in drawing conclusions due to the uncertainty in determining the sources of the variations in the results. Clearly, assessor differences or apple heterogeneity could be responsible for these variations. Moreover, according to Hampson et al. (2000) who studied genotype differences from a sensory point of view, apple heterogeneity may cause difficulty in differentiating samples. In fact, real variations within a given genotype may make differences among genotypes harder to detect.

To study sample differences in sensory evaluation, a version of a mixed model analysis of variance (ANOVA) is commonly performed for each attribute. ANOVA models have been discussed and debated to take into account the particular nature of sensory data better, such as assessor differences, replicates, etc. Despite the development of specific analyses that meet the special requirements of sensory data, such as the assessor model (Brockhoff, 2003; Brockhoff & Skovgaard, 1994) and its extended version (Brockhoff, Schlich, & Skovgaard, 2013), the standard model is most often used. This analysis includes a fixed sample effect, a random assessor effect and the interaction between sample and assessor. This model is applied in order to obtain information about the samples while accounting properly for possible assessor differences. However, the interaction term accounts for both disagreement and scaling differences and is generally falsely interpreted as only disagreement. Moreover, in statistical data analysis, it is good practice to model all the different effects involved. So, regarding sensory data, the scaling effect and the unit effect, in the

* Corresponding author. Address: GROUPE ESA, Unité de Recherche GRAPPE, Ecole Supérieure d'Agriculture d'Angers' 55 Rue Rabelais, BP 30748, F-49007 Angers Cedex 01, France. Tel.: +33 2 41 23 55 55; fax: +33 2 41 23 55 00.
E-mail address: c.bavay@groupe-esa.com (C. Bavay).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.06.003

specific case of a heterogeneous product, should be included in the analysis.

The focus of the present paper is the modeling of the variability of the sensory response in descriptive sensory analysis in order to understand the results better. Data are analyzed using three models: the standard model, a model considering sample heterogeneity and the extended mixed assessor model approach. The contributions of the model considering heterogeneity are investigated in comparison with the standard model. Then, the contributions of the extended assessor model approach, in combination with accounting for the complex replication structure, are studied. Inclusion of within-sample heterogeneity may, with good reason, affect other noise parts of the model, e.g. the important assessor-by-sample interaction. The main goal of the scaling correction, which comes from using the assessor model approach, is to affect this interaction. We demonstrate the possibility and the advantage of using the assessor model approach in combination with the inclusion of other effects, such as within-sample heterogeneity. Using the assessor model in combination with accounting for this more complex and unbalanced sample replication structure is novel.

2. Materials and methods

2.1. Models

Model analyses are run with the step function of the lmerTest package (Kuznetsova, Christensen, Bavay, & Brockhoff, 2013; Kuznetsova, Christensen, & Brockhoff, 2012) using the R software (version 2.14.2) (R Core Team, 2012). The *F* test and log likelihood ratio test are applied to test for fixed effects and for random effects, respectively. R codes are provided in the appendix for useRs.

2.1.1. Standard model

The standard model for analyzing sensory data is:

$$X_{asr} = \upsilon_s + \alpha_a + \gamma_{as} + \varepsilon_{asr} \quad (1)$$

where $\alpha_a \sim N(0, \sigma^2_{assessor})$, $\gamma_{as} \sim N(0, \sigma^2_{sample*assessor})$ and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

The model includes a fixed sample effect υ_s, a random assessor effect α_s and the interaction between sample and assessor γ_{as}. The *r* subscript accounts for random replicates. This model is applied in order to obtain information about the samples while accounting properly for possible assessor differences. The assessor effect accounts for level differences while the interaction term accounts for both disagreement and scaling differences.

2.1.2. Model considering within-sample heterogeneity

Consider a sensory experiment with no session effect and a simple one-way sample structure. Each sample is made up of several units (e.g. individual fruits within an apple cultivar). These units may present differences and we therefore want to take into account the main unit effect. With that aim, a random unit effect nested within the sample effect $\delta_{u(asr)}$ (e.g. a fruit nested within an apple cultivar) is introduced into the model:

$$X_{asr} = \upsilon_s + \alpha_a + \gamma_{as} + \delta_{u(asr)} + \varepsilon_{asr} \quad (2)$$

where $\alpha_a \sim N(0, \sigma^2_{assessor})$, $\gamma_{as} \sim N(0, \sigma^2_{assessor*sample})$, $\delta_{u(asr)} \sim N(0, \sigma^2_{unit})$ and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

As in model (1), the *r* subscript accounts for replications of the measurement. In our example, replicates consist of pieces of apple (apples are units and each apple is cut into pieces). In the subscript *u*(*asr*), *u* accounts for the actual unit (a fruit) and *asr* indicates the numbering of the actual observation.

Model (2) takes into account assessor level differences, assessor scaling differences and disagreement included in the interaction term and actual unit differences. The introduction of the unit effect is made possible by having a single unit rated by several assessors.

2.1.3. Assessor model approach

The original assessor model for sensory data was proposed by Brockhoff and Skovgaard (1994) and only included the scaling differences (as fixed effects):

$$X_{asr} = \alpha_a + \upsilon_s \cdot \beta_a + \varepsilon_{asr} \quad (3)$$

where $\varepsilon_{asr} \sim N(0, \sigma^2_a)$; all terms are independent.

The model comprises a fixed sample effect υ_s, a fixed assessor effect α_a and the individual scaling coefficient β_a.

In Brockhoff (2003), this model together with other different models were further developed into an approach for univariate assessor performance investigations. In Brockhoff et al. (2013), an extended version of the assessor model including a random interaction (disagreement) effect was presented using the centered sample means m_s as covariates:

$$X_{asr} = \alpha_a + \upsilon_s + m_s \cdot \beta_a + \gamma_{as} + \varepsilon_{asr} \quad (4)$$

where $\gamma_{as} \sim N(0, \sigma^2_{assessor*sample})$, and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

This model includes a fixed sample effect υ_s, a random assessor effect α_s, the interaction between sample and assessor γ_{as}, the individual scaling coefficient β_a and the centered sample means m_s. In their paper, the authors show how proper hypothesis testing for sample comparisons can be based on this model, which amounts to simply removing the scaling part of the interaction by linear regression (although proper confidence bands would require more complicated computations).

To investigate the consequence of both the scaling/disagreement decomposition and the unit effect, we would like to extend the model in (5) with the random unit effect:

$$X_{asr} = \alpha_a + \upsilon_s + m_s \cdot \beta_a + \gamma_{as} + \delta_{u(asr)} + \varepsilon_{asr} \quad (5)$$

where $\gamma_{as} \sim N(0, \sigma^2_{assessor*sample})$, $\delta_{u(asr)} \sim N(0, \sigma^2_{unit})$ and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

It is beyond the scope of the present paper to provide a full methodological treatment of this model (5) applied to unbalanced data, which has not been presented in the literature with its extension. We apply here a simple approach to investigate both effects. We construct a processed version of the data where we have removed (additively) the scaling part of the interaction, similar to the "additive approach" suggested and discussed in Romano et al. (2008). Here, this is done by applying the version of the assessor model (4), where the scaling effects are estimated based on the centered sample means m_s and then subtracted from the data:

$$X_{asr} - m_s \cdot (\beta_a - \overline{\beta}) \quad (6)$$

2.1.4. Random component models

To study the relative sizes of the various effects, a version of models (1) and (2) above, where all effects are considered random (also the sample effect), is applied:

$$X_{asr} = \upsilon_s + \alpha_a + \gamma_{as} + \varepsilon_{asr} \quad (7)$$

where $\upsilon_s \sim N(0, \sigma^2_{sample})$, $\alpha_a \sim N(0, \sigma^2_{assessor})$, $\gamma_{as} \sim N(0, \sigma^2_{assessor*sample})$ and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

$$X_{asr} = \upsilon_s + \alpha_a + \gamma_{as} + \delta_{u(asr)} + \varepsilon_{asr} \quad (8)$$

where $\upsilon_s \sim N(0, \sigma^2_{sample})$, $\alpha_a \sim N(0, \sigma^2_{assessor})$, $\gamma_{as} \sim N(0, \sigma^2_{assessor*sample})$, $\delta_{u(asr)} \sim N(0, \sigma^2_{unit})$ and $\varepsilon_{asr} \sim N(0, \sigma^2)$; all terms are independent.

The last model, including the unit effect, is then applied to both the original data and the data with the scaling removed. For the

original data analyzed by the simple model (7), the variances are then:

$$\mathrm{Var}(X) = \sigma^2_{sample} + \sigma^2_{assessor} + \sigma^2_{assessor*sample} + \sigma^2 \quad (9)$$

Inclusion of the unit effect provides the following:

$$\mathrm{Var}(X) = \sigma^2_{sample} + \sigma^2_{assessor} + \sigma^2_{assessor*sample} + \sigma^2_{unit} + \sigma^2 \quad (10)$$

Finally, for the scaling-corrected data, this becomes:

$$\mathrm{Var}(X) = \sigma^2_{sample} + \sigma^2_{assessor} + \sigma^2_{disagreement} + \sigma^2_{unit} + \sigma^2 \quad (11)$$

2.2. Data

The data come from a sensory study of apples. Three different cultivars were tested: Ariane, Braeburn and Pink Lady®. There were 19 assessors in total and they tested each apple cultivar four times. For testing, each apple was cut into pieces and distributed to three or four assessors so that 18 apples were tested for each cultivar. The assessors were asked to score seven sensory attributes, namely *crunchiness*, *firmness*, *crispness*, *juiciness*, *fondant*, *acidity* and *sweetness*. To avoid confusion, from now on the term sample will be used to refer to the cultivar and the term unit will be used to refer to the individual fruit.

3. Results and discussion

3.1. From the standard model to the model accounting for within-sample heterogeneity

For each attribute, a model with and without the unit effect (models (2) and (1), respectively) was fitted together with the variance component versions (7) and (8). The variance components for the seven attributes for the standard model and the model that takes into account within-sample heterogeneity are displayed in Figs. 1 and 2, respectively. For each attribute except *sweetness*, the estimated part of the variation of the unit effect is large, always representing more than 10% of the total variance. These results indicate that there are large differences between individual fruits within each cultivar. A major change is observed for the residual

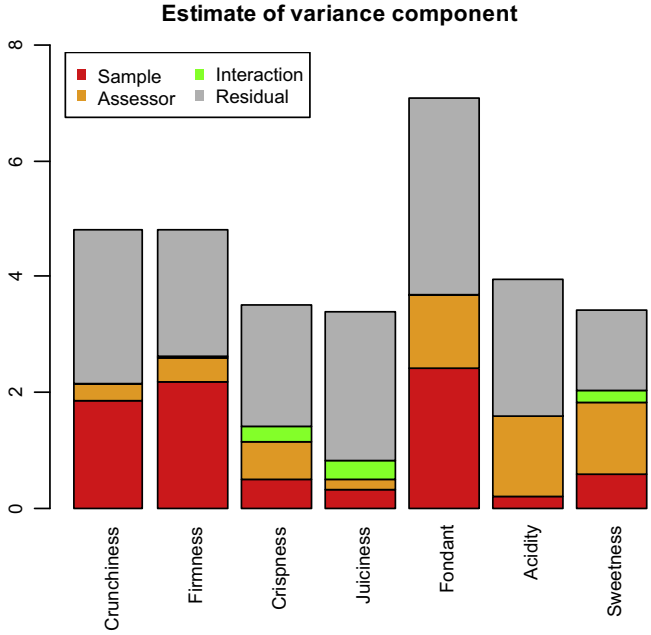


Fig. 1. Barplots for estimate of variance component for standard model.

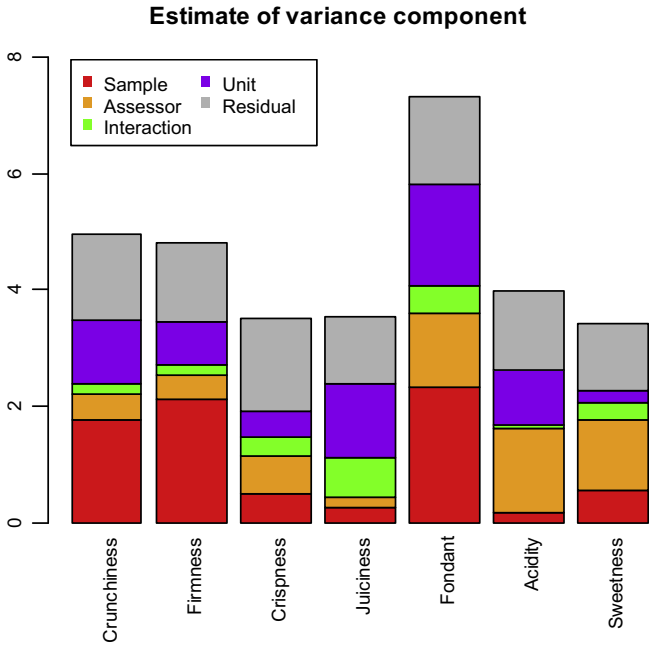


Fig. 2. Barplots for estimate of variance component for model accounting for intra-batch variability.

part of the variance. The addition of the unit effect causes the residual part of the variance observed for model (1) to be split into the unit part and the residual part for the results of model (2). The sample and assessor parts of the variance do not change much. It should be noticed that the addition of the unit effect leads to an increase in the part of the variance attributed to the assessor-by-sample interaction. This phenomenon may be due to the decrease in the part of the variance attributed to error. In fact, the residual part of the variance plays a role in the variance component computation. In unbalanced data and partly "crossed" cases like this, the inclusion of a relevant variability source, like the unit effect, may simply lead to a more correct estimation of all other effects. Such changes in other effects can generally go in either direction: other effects may lose or gain importance.

The statistics as well as the corresponding p-values are displayed in Table 1. In the results for model (2), the effect of the unit factor is significant for all attributes except *sweetness*. F statistics of the sample effect decrease when the unit effect is added to the model. This phenomenon is explained by an increase in the noise in the *F* ratio. In fact, in model (2), the noise part of the *F* ratio calculated for the sample effect is more complex than the simple assessor-by-sample interaction term (used in model (1)). In the present case, the noise part includes the unit. The level of significance of the sample effect is unchanged for *crunchiness*, *firmness*, *crispness*, *fondant* and *sweetness*. For *juiciness* and *acidity*, the level of significance is lower in the case of model (2). For these two attributes, the part of the variance due to the unit effect is more than five times larger than the variance part due to the sample effect. It should be noticed that, with model (1), the level of significance of these two attributes is lower than for the other attributes. Here, the importance of accounting for the within-sample heterogeneity is illustrated. Indeed, conclusions about the differences perceived between the samples differ according to the model. In the case of model (2), with a 95% level of confidence, samples were not declared different for *juiciness* (only a trend could be declared). For *acidity*, in both cases the samples were declared different but with 95% confidence for model (2) against 99% confidence for model (1). Regarding the F statistics for the interaction term, it generally

Table 1
Statistics (*F* value for the sample and scaling effects and Chi^2 for random effects) and *p* values for the three models.

Descriptor	Factor	Standard model		Model accounting for biological variability		Mixed assessor model accounting for biological variability	
		Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
Crunchiness	Cultivar	40.25	1.05E−14	21.01	2.89E−07	23.28	9.03E−08
	Assessor	4.83	2.79E−02	15.49	8.30E−05	19.83	8.46E−06
	Interaction[a]	0.00	9.99E−01	0.68	4.09E−01	0.00	9.98E−01
	Scaling	–	–	–	–	0.94	5.52E−01
	Fruit	–	–	20.05	7.53E−06	22.1	2.59E−06
Firmness	Cultivar	57.59	2.76E−19	33.61	1.11E−09	35.32	5.34E−10
	Assessor	10.52	1.18E−03	13.76	2.08E−04	20.84	4.98E−06
	Interaction	0.00	1.00E+00	1.19	2.74E−01	0.00	9.96E−01
	Scaling	–	–	–	–	1.28	3.01E−01
	Fruit	–	–	11.43	7.24E−04	18.65	1.57E−05
Cripsness	Cultivar	13.86	3.00E−06	13.86	3.00E−06	16.65	2.95E−07
	Assessor	20.57	5.74E−06	20.57	5.74E−06	27.65	1.45E−07
	Interaction	1.06	3.04E−01	2.45	1.18E−01	0.00	9.98E−01
	Scaling	–	–	–	–	1.23	3.34E−01
	Fruit	–	–	3.4	6.52E−02	3.34	6.78E−02
Juiciness	Cultivar	7.17	1.02E−03	2.84	6.54E−02	4.26	1.98E−02
	Assessor	0.77	3.80E−01	0.77	3.82E−01	12.39	4.32E−04
	Interaction	3.75	5.28E−02	18.92	1.37E−05	3.02	8.20E−02
	Scaling	–	–	–	–	0.50	9.22E−01
	Fruit	–	–	29.62	5.25E−08	26.43	2.74E−07
Fondant	Cultivar	41.39	4.85E−15	16.56	3.18E−06	20.94	3.27E−07
	Assessor	24.93	5.93E−07	14.49	1.41E−04	51.92	5.79E−13
	Interaction	0.00	9.96E−01	4.95	2.60E−02	0.05	8.23E−01
	Scaling	–	–	–	–	0.84	6.43E−01
	Fruit	–	–	29.91	4.53E−08	33.27	8.03E−09
Acidity	Cultivar	5.97	3.20E−03	3.26	4.71E−02	3.23	4.84E−02
	Assessor	42.72	6.31E−11	58.73	1.81E−14	71.84	2.33E−17
	Interaction	0.00	1.00E+00	0.14	7.11E−01	0.00	9.97E−01
	Scaling	–	–	–	–	1.05	4.62E−01
	Fruit	–	–	22.34	2.29E−06	31.61	1.89E−08
Sweetness	Cultivar	22.63	2.58E−09	22.63	2.58E−09	26.02	1.99E−10
	Assessor	60.68	6.66E−15	60.68	6.66E−15	70.42	0.00E+00
	Interaction	1.76	1.85E−01	1.76	1.85E−01	0.00	1.00E+00
	Scaling	–	–	–	–	0.68	7.85E−01
	Fruit	–	–	2.22	1.37E−01	1.84	1.74E−01

[a] Interaction corresponds to [Disagreement + Scaling] in the standard model and in the model accounting for biological variability and to Disagreement only in the mixed assessor model accounting for biological variability.

increases from model (1) to model (2). The significance test consists of the relationship between the interaction mean square and the residual. So, the large change in the residual explains the change in the interaction term.

As a conclusion, not accounting for unit variations (individual fruit differences in our example) may lead to an incorrect interpretation of the results. Mistaken conclusions could be drawn about sample differences.

3.2. From the model accounting for within-sample heterogeneity to the general mixed assessor model approach

Now within-sample heterogeneity is accounted for, we would like a better analysis of assessor differences, especially the assessor-by-sample interaction part. Data are corrected for the scaling effect, therefore applying model (2) will result in an estimation of disagreement as the interaction term. The results of model (2) applied to corrected data are compared with the results of model (2) applied to raw data and, similarly, the random version model (8) is applied to both raw and corrected data. Fig. 3 shows the estimates of variance components from the analyses of corrected data. Comparing with Fig. 2, it is clear that the interaction variance seen in Fig. 2, even though already quite small, has almost completely disappeared in Fig. 3. It even becomes zero (or almost zero) for all attributes but *juiciness*. These results show that the interaction term seen in Fig. 2, with the combined interpretation as both

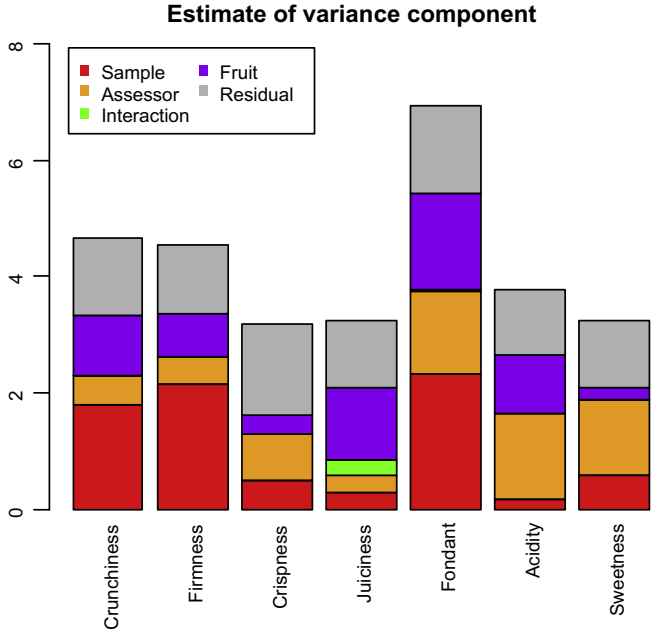


Fig. 3. Barplots for estimate of variance component for mixed assessor model accounting for intra-batch variability.

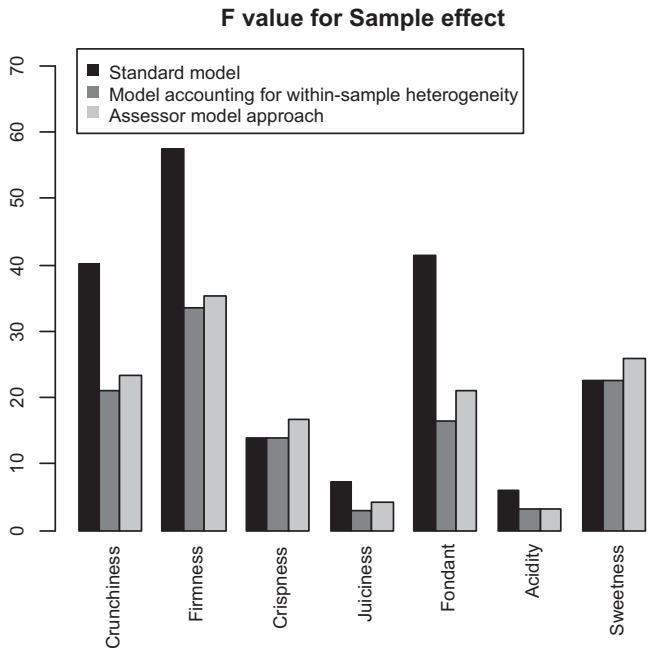


Fig. 4. Barplots for *F* value for Sample effect for the three models.

disagreement and scaling differences, appeared to be almost entirely a scaling effect for these data, which has been removed from the results in Fig. 3. Removing the scaling effect before the analysis does not lead to large changes in the parts of the variance contributed by the sample, the assessor, the fruit and error.

Regarding the significance test, there is a slight increase in the sample effect significance (see Fig. 4). This is explained by the reduction in the noise part in the *F* ratio through the scaling correction. The significance of the assessor effect also increases, probably because of the slight decrease in the error. The unit effect is not much affected. Considering the interaction term or disagreement, it is non-significant. The significance of the scaling effect was also computed and found to be not really significant. This is due to the generally very low interaction level of these data.

4. Conclusion

Significance may be declared in the case of the standard model, whereas including the unit effect leads to a less clear conclusion (e.g. *juiciness*). Applying the right model is therefore crucial, in order to obtain reliable and repeatable results.

The results concerning interaction vary with the model. However, these results primarily show little interaction. Such a model should be applied to data presenting significant interaction (from model (1)) and then compared to the results from the improved model. Correcting for scale use will have a larger impact for data exhibiting more interaction.

In conclusion, the two issues raised in this paper, inclusion of unit effects and correcting for scaling differences, have been shown, as generally expected, to pull in opposite directions regarding the sensitivity of the study to find product differences. Omitting the unit effect can lead to major over-optimism in the significance findings related to product differences. Omitting the scaling correction can lead to the opposite outcome – the extent of which depends on the level of interaction in general. In addition, for both effects, their inclusion in the modeling leads to greater understanding of the important sources of variability in data and thus of the situation being studied. This knowledge gained about variability will help in designing future experiments.

Acknowledgements

The authors thank Corinne Patron and Isabel Saillard for their assistance, and the assessors who participated in the descriptive sensory analysis. This research was funded by the Conseil Régional des Pays de Loire.

Appendix A

```
# Package loading
library(lme4)
library(lmerTest)
# For a given attribute in column i
### Analysis of mixed models
### Standard model
res = lmer(donnee[,i]~Sample+(1|Assessor)+(1|Assessor:Sam ple),data=donnee)
tres = step(res, reduce.random=FALSE, reduce.fixed=FALSE)
# Displaying results for fixed effect
print(anova.table)
# Displaying results for random effect
print(rand.table)
### Model considering within-sample heterogeneity
res = lmer(donnee[,i]~Sample+(1|Assessor)+(1|Assessor:Sam ple)+(1|Unit),data=donnee)
tres = step(res, reduce.random=FALSE, reduce.fixed=FALSE)
# Displaying results for fixed effect
print(anova.table)
# Displaying results for random effect
print(rand.table)
### Assessor model approach of the model considering within-sample heterogeneity
x.temp<-tapply(donnee[,i],donnee$Sample, mean)
donnee$X<-unlist(lapply(donnee$Sample, function(x) return(x.temp[names(x.temp) %in% x] - mean(donnee[,i]))))
res.lm.donnee<-residuals(lm(donnee[,i]~Sample+Assessor:X, donnee))
donnee$newResp<-unlist(lapply(donnee$Sample, function(x) return(x.temp[names(x.temp) %in% x]))) + res.lm.donnee
res = lmer(newResp~Sample+(1|Assessor)+(1|Assessor:Sam ple)+(1|Unit),data=donnee)
tres = step(res, reduce.random=FALSE, reduce.fixed=FALSE)
# Displaying results for fixed effect
print(anova.table)
# Displaying results for random effect
print(rand.table)
#### Random component models
#### Standard model
res = lmer(donnee[,i]~(1|Sample)+(1|Assessor)+(1|Assessor:Sample),data=donnee)
# the object REmat contains the values for variance and standard deviation for each random effect
var<-summary(res)@REmat
# Plotting the variance component as a cumulative barplot
barplot(as.matrix(var[,3]), legend.text=var[,1])
### Model considering within-sample heterogeneity
res = lmer(donnee[,i]~(1|Sample)+(1|Assessor)+(1|Assessor:Sample)+(1|Unit),data=donnee)
# the object REmat contains the values for variance and standard deviation for each random effect
var<-summary(res)@REmat
# Plotting the variance component as a cumulative barplot
barplot(as.matrix(var[,3]), legend.text=var[,1])
### Assessor model approach of the model considering within-sample heterogeneity
x.temp<-tapply(donnee[,i],donnee$Sample, mean)
```

```
donnee$X<-unlist(lapply(donnee$Sample, function(x) return (x.temp[names(x.temp) %in% x] - mean(donnee[,i]))))
res.lm.donnee<-residuals(lm(donnee[,i]~Sample+Assessor:X, donnee))
donnee$newResp<-unlist(lapply(donnee$Sample, function(x) return(x.temp[names(x.temp) %in% x]))) + res.lm.donnee
res = lmer(newResp~(1|Sample)+(1|Assessor)+(1|Asses sor: Sample)+(1|Unit),data=donnee)
# the object REmat contains the values for variance and standard deviation for each random effect
var<-summary(res)@REmat
# Plotting the variance component as a cumulative barplot
barplot(as.matrix(var[,3]), legend.text=var[,1])
```

References

Brockhoff, P. B. (1998). Assessor modelling. *Food Quality and Preference, 9*(3), 87–89.

Brockhoff, P. B. (2003). Statistical testing of individual differences in sensory profiling. *Food Quality and Preference, 14*(5–6), 425–434.

Brockhoff, P. B., Schlich, P., & Skovgaard, I. M. (2013). Accounting for scaling differences in sensory profile data: Improved mixed model analysis of variance. *Manuscript Intended for Food Quality and Preference*.

Brockhoff, P. B., & Skovgaard, I. M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference, 5*(3), 215–224.

Hampson, C. R., Quamme, H. A., Hall, J. W., MacDonald, R. A., King, M. C., & Cliff, M. A. (2000). Sensory evaluation as a selection tool in apple breeding. *Euphytica, 111*(2), 79–90.

Kuznetsova, A., Christensen, R. H. B., Bavay, C., & Brockhoff, P. B. (2013). Automized mixed ANOVA modelling of sensory and consumer data. *Manuscript Intended for Food Quality and Preference*.

Kuznetsova, A., Christensen, R.H.B., & Brockhoff, P.B. (2012). *lmerTest: Tests for random and fixed effects for linear mixed effect models* (lmer objects of lme4 package). R package version 1.0-2.

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference, 2*(3), 187–199.

Næs, T. (1998). Detecting individual differences among assessors and differences among replicates in sensory profiling. *Food Quality and Preference, 9*(3), 107–110.

Næs, T., & Solheim, R. (1991). Detection and interpretation of variation within and between assessors in sensory profiling. *Journal of Sensory Studies, 6*(3), 159–177.

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romano, R., Brockhoff, P. B., Hersleth, M., Tomic, O., & Næs, T. (2008). Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. *Food Quality and Preference, 19*(2), 197–209.

Schlich, P. (1994). Grapes: A method and a SAS® program for graphical representations of assessor performances. *Journal of Sensory Studies, 9*(2), 157–169.

Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. *Data Handling in Science and Technology, 16*, 259–306.

Williams, A. A., & Carter, C. S. (1977). A language and procedure for the sensory assessment of Cox's Orange Pippin apples. *Journal of the Science of Food and Agriculture, 28*(12), 1090–1104.

Appendix G

Reference manual for the R package lmerTest

Alexandra Kuznetsova, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen (2013) Reference manual for the R package SensMixed

Package 'lmerTest'

December 23, 2014

Type Package

Title Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).

Version 2.0-22

Date 2012-01-09

Maintainer Alexandra Kuznetsova <alku@dtu.dk>

Depends R (>= 3.0.0), Matrix, stats, methods, lme4 (>= 1.0)

Imports plyr, numDeriv, MASS, Hmisc, gplots

Suggests pbkrtest

Description The package provides different kinds of tests on lmer objects (of lme4 package). The tests comprise type 3 and type 1 F tests for fixed effects, LRT tests for random effects, calculation of population means for fixed factors with confidence intervals and corresponding plots. Package also provides backward elimination of non-significant effects

LazyData TRUE

License GPL (>= 2)

Repository CRAN

Date/Publication 2013-01-26 08:14:39

Author Alexandra Kuznetsova [aut, cre],
Per Bruun Brockhoff [aut, ths],
Rune Haubo Bojesen Christensen [aut]

R topics documented:

`lmerTest-package`	*The package performs different kinds of tests on lmer objects, such as* F *tests of type 3/type 1 hypotheses for the fixed part, likelihood ratio tests for the random part, least squares means (population means) and differences of least squares means for the factors of the fixed part with corresponding plots. The package also provides with a function step, that preforms backward elimination of non-significant effects, starting from the random effects, and then fixed ones.*

Description

The package provides anova function, that gives data frame similar to what gives **lme4** package but with p-values calculated from F statistics of type 3/type 1 hypotheses. There are two options for denominator degrees of freedom of F statistics: "Satterthwaite" and "Kenward-Roger". The calculation of anova with Kenward-Roger's approximation is based on function from **pbkrtest** package, the calculation of Satterthwaite's approximation is based on SAS proc mixed theory (see reference).The type 3 hypothesis (marginal) is calculated according to SAS theory (SAS Institute Inc., 1978). The package also provides `summary` function, which gives the same as **lme4** package summary function but with p-values and degrees of freedom added for the t-test (based on Satterthwaite approximation for denominator degrees of freedom). The tests on random effects are performed using likelihood ratio tests.

Details

Package:	lmerTest
Type:	Package
Version:	1.0
Date:	2012-01-10
License:	GPL

The calculation of statistics for the fixed part was developed according to SAS Proc Mixed Theory (see reference).

Author(s)

Alexandra Kuznetsova <alku@dtu.dk>, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

References

SAS Technical Report R-101 1978 Tests of Hypotheses in Fixed-Effects Linear Models *Copyright (C)* (SAS Institute Inc., Cary, NC, USA)

Goodnight, J.H. 1976 General Linear Models Procedure (S.A.S. Institute, Inc.)

Schaalje G.B., McBride J.B., Fellingham G.W. 2002 Adequacy of approximations to distributions of test Statistics in complex mixed linear models

Examples

```
#import lmerTest package
```

```
library(lmerTest)

# an object of class merModLmerTest
m <- lmer(Informed.liking ~ Gender+Information+Product +(1|Consumer), data=ham)

# gives summary of lmer object. The same as of class merMod but with
# additional p-values calculated based on Satterthwates approximations
summary(m)

# anova table the same as of class merMod but with additional F statistics and
#  and denominator degrees of freedom and
# p-values calculated based on Satterthwaites approximations
anova(m)

# anova table the same as of class merMod but with additional F statistics and
# denominator degrees of freedom and
# p-values calculated based on Kenward-Rogers approximations
## Not run:
if(require(pbkrtest))
anova(m, ddf = "Kenward-Roger")

## End(Not run)

# anova table of class merMod
anova(m, ddf="lme4")

# backward elimination of non-significant effects of model m
st <- step(m)

plot(st)
```

---

anova-methods *Methods for function* anova *in package* **lmerTest**

Description

Methods for Function anova in Package **lmerTest**

Usage

```
## S4 method for signature merModLmerTest
anova(object, ... , ddf="Satterthwaite",
type=3)
```

Arguments

object	object of class "merModLmerTest"
...	object of class "merModLmerTest". Then the model comparison statistisc will be calculated
ddf	By default the Satterthwaite's approximation to degrees of freedom is calculated. If ddf="Kenward-Roger", then the Kenward-Roger's approximation is calculated using KRmodcomp function from **pbkrtest** package. If ddf="lme4" then the anova table that comes from **lme4** package is returned.

type	type of hypothesis to be tested. Could be type=3 or type=1 (The definition comes from SAS theory)

References

SAS Technical Report R-101 1978 Tests of Hypotheses in Fixed-Effects Linear Models *Copyright (C)* (SAS Institute Inc., Cary, NC, USA)

Goodnight, J.H. 1976 General Linear Models Procedure (S.A.S. Institute, Inc.)

Schaalje G.B., McBride J.B., Fellingham G.W. 2002 Adequacy of approximations to distributions of test Statistics in complex mixed linear models

Examples

```
#import lmerTest package
library(lmerTest)

m.ham <- lmer(Informed.liking ~ Product*Information*Gender
+ (1|Consumer), data = ham)

# type 3 anova table with  denominator degrees of freedom
# calculated based on Satterthwaites approximation
anova(m.ham)

# type 1 anova table with denominator degrees of freedom
# calculated based on Satterthwaites approximation
## Not run:
anova(m.ham, type = 1)

## End(Not run)

# type3 anova table with additional F statistics and denominator degrees of freedom
# calculated based on Kenward-Rogers approximation
if(require(pbkrtest))
anova(m.ham, ddf = "Kenward-Roger")

## Not run:
# anova table, that is returned by lme4 package
anova(m.ham, ddf = "lme4")

## End(Not run)
```

---

carrots *Consumer preference mapping of carrots*

Description

In a consumer study 103 consumers scored their preference of 12 danish carrot types on a scale from 1 to 7. Moreover the consumers scored the degree of sweetness, bitterness and crispiness in the products. The carrots were harvested in autumn 1996 and tested in march 1997. In addition to the consumer survey, the carrot products were evaluated by a trained panel of tasters, the sensory panel, with respect to a number of sensory (taste, odour and texture) properties. Since usually a high

number of (correlated) properties(variables) are used, in this case 14, it is a common procedure to use a few, often 2, combined variables that contain as much of the information in the sensory variables as possible. This is achieved by extracting the first two principal components in a principal components analysis(PCA) on the product-by-property panel average data matrix. In this data set the variables for the first two principal components are named (`sens1` and `sens2`).

Usage

```
carrots
```

Format

`Consumer` factor with 103 levels: numbering identifying consumers

`Frequency` factor with 5 levels; "How often do you eat carrots?" 1: once a week or more, 2: once every two weeks, 3: once every three weeks, 4: at least once month, 5: less than once a month

`Gender` factor with 2 levels. 1: male, 2:female

`Age` factor with 4 levels. 1: less than 25 years, 2: 26-40 years, 3: 41-60 years, 4 more than 61 years

`Homesize` factor with two levels. Number of persons in the household. 1: 1 or 2 persons, 2: 3 or more persons

`Work` factor with 7 levels. different types of employment. 1: unskilled worker(no education), 2: skilled worker(with education), 3: office worker, 4: housewife (or man), 5: independent businessman/ self-employment, 6: student, 7: retired

`Income` factor with 4 levels. 1: <150000, 2: 150000-300000, 3: 300000-500000, 4: >500000

Source

Per Bruun Brockhoff, The Royal Veterinary and Agricultural University, Denmark.

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)

m.carrots <- lmer(Preference ~ sens2 + Homesize
+(1+sens2|Consumer), data=carrots)

# only elimination of the random part is required.
#approximation of ddf is Satterthwaite
step(m.carrots, reduce.random = FALSE)
```

`difflsmeans` *Calculates Differences of Least Squares Means and Confidence Intervals for the factors of a fixed part of mixed effects model of lmer object.*

Description

Produces a data frame which resembles to what SAS software gives in proc mixed statement. The approximation for degrees of freedom is Satterthwaite's.

Usage

```
difflsmeans(model, test.effs=NULL,  ...)
```

Arguments

model	linear mixed effects model (lmer object).
test.effs	charachter vector specyfying the names of terms to be tested. If NULL all the terms are tested.
...	other potential arguments.

Value

Produces Differences of Least Squares Means (population means) table with p-values and Confidence intervals.

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

See Also

lsmeans, step, rand

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)

#specify lmer model
m1 <- lmer(Informed.liking ~ Gender*Information +(1|Consumer), data=ham)

#calculate least squares means for interaction Gender:Information
difflsmeans(m1, test.effs="Gender:Information")

#import TVbo data from lmerTest package
data(TVbo)

m <- lmer(Coloursaturation ~ TVset*Picture + (1|Assessor), data=TVbo)
plot(difflsmeans(m, test.effs="TVset"))
```

---

ham *Conjoint study of dry cured ham*

Description

One of the purposes of the study was to investigate the effect of information given to the consumers measured in hedonic liking for the hams. Two of the hams were Spanish and two were Norwegian, each origin representing different salt levels and different aging time. The information about origin was given in such way that both true and false information was given. essentially a 4*2 design with 4 samples and 2 information levels. A total of 81 Consumers participated in the study.

Usage

```
ham
```

Format

`Consumer` factor with 81 levels: numbering identifying consumers

`Product` factor with four levels

`Informed.liking` numeric: hedonic liking for the products

`Information` factor with two levels

`Gender` factor with two levels (gender)

`Age` numeric: age of Consumer

References

"Alternative methods for combining design variables and consumer preference with information about attitudes and demographics in conjoint analysis" . T. Naes, V.Lengard, S. Bolling Johansen, M. Hersleth

Examples

```
#import lmerTest package
library(lmerTest)

m <- lmer(Informed.liking ~ Product*Information*Gender
+ (1|Product:Consumer) , data=ham)

#anova table with p-values with Satterthwaites approximation for denominator
#degrees of freedom
anova(m)

#analysis of random and fixed parts and post hoc
#analysis of Product and Information effects
step(m, reduce.random=FALSE, reduce.fixed=FALSE,
test.effs=c("Product", "Information"))
```

`lmer` *Fit Linear Mixed-Effects Models*

Description

Fit a linear mixed model

Details

This `lmer` function is an overloaded function of `lmer` (`merMod` class from **lme4** package).

Value

An object of class `"merModLmerTest"`

See Also

merModLmerTest class

Examples

```
library(lmerTest)

## linear mixed models
fm1 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy)
fm2 <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), sleepstudy)

# anova table the same as of class merMod but with additional F statistics and
# p-values calculated based on Satterthwaites approximations
anova(fm1)

# anova table the same as of class merMod but with additional F statistics and
# p-values calculated based on Kenward-Rogers approximations
## Not run:
if(require(pbkrtest))
anova(fm1, ddf="Kenward-Roger")

# anova table the same as of class merMod
anova(fm1, ddf="lme4")

## End(Not run)

# gives summary of merModLmerTest class. The same as of class merMod but with
# additional p-values calculated based on Satterthwates approximations
summary(fm1)

## multiple comparisons statistics. The one from lme4 package
## Not run:
anova(fm1, fm2)

## End(Not run)
```

`lsmeans` *Calculates Least Squares Means and Confidence Intervals for the factors of a fixed part of mixed effects model of lmer object.*

Description

Produces a data frame which resembles to what SAS software gives in proc mixed statement. The approximation of degrees of freedom is Satterthwate's.

Usage

```
lsmeans(model, test.effs=NULL,  ...)
```

Arguments

model	linear mixed effects model (lmer object).
test.effs	charachter vector specyfying the names of terms to be tested. If NULL all the terms are tested.
...	other potential arguments.

Value

Produces Least Squares Means (population means) table with p-values and Confidence intervals.

Note

For construction of the contrast matrix popMatrix function from **doBy** package was used.

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

References

doBy package, **gplots** package

See Also

step, rand, difflsmeans

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)

#specify lmer model
m1 <- lmer(Informed.liking ~ Gender*Information +(1|Consumer), data=ham)

#calculate least squares means for interaction Gender:Information
lsmeans(m1, test.effs="Gender:Information")

#import TVbo data from lmerTest package
data(TVbo)

m <- lmer(Coloursaturation ~ TVset*Picture + (1|Assessor), data=TVbo)
plot(lsmeans(m))
lsmeans(m, test.effs="TVset")
```

merModLmerTest-class *Mixed Model Representations*

Description

The merModLmerTest *contains* merMod class of **lme4** package and overloads anova and summary functions.

Objects from the Class

Objects can be created via the lmer functions.

See Also

lmer()

Examples

```
(m <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject),
             data = sleepstudy))

# type 3 anova table with  denominator degrees of freedom
# calculated based on Satterthwaites approximation
anova(m)

# type 1 anova table with denominator degrees of freedom
# calculated based on Satterthwaites approximation
## Not run:
anova(m, type=1)

## End(Not run)

# type3 anova table with additional F statistics and denominator degrees of freedom
# calculated based on Kenward-Rogers approximation
## Not run:
if(require(pbkrtest))
anova(m, ddf="Kenward-Roger")

## End(Not run)

# anova table, that is returned by lme4 package
anova(m, ddf="lme4")

# summary of merModLmerTest object. Returns the same as merMod object but with an
#additional column of p values for the t test.
summary(m)
```

rand *Performs likelihood ratio test on random effects of linear mixed effects model.*

Description

Returns a data frame with values of Chi square statistics and corresponding p-values of likelihood ratio tests.

Usage

```
rand(model, ...)
```

Arguments

model	linear mixed effects model (lmer object).
...	other potential arguments.

Details

The columns of the data are:

Chisq: The value of the chi square statistics

Chi Df: The degrees of freedom for the test

p.value: The p-value of the likelihood ratio test for the effect

Value

Produces a data frame with LR tests for the random terms.

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

See Also

step, lsmeans, difflsmeans

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)

#lmer model with correlation between intercept and slopes
#in the random part
m <- lmer(Preference ~ sens2+Homesize+(1+sens2|Consumer), data=carrots)

# table with  p-values for the random effects
rand(m)
```

step *Performs backward elimination of non-significant effects of linear mixed effects model:*

Description

performs automatic backward elimination of all effects of linear mixed effect model. First backward elimination of the random part is performed following by backward elimination of the fixed part. Finally LSMEANS (population means) and differences of LSMEANS for the fixed part of the model are calculated and the final model is provided. The p-values for the fixed effects are calculated from F test based on Sattethwaite's or Kenward-Roger approximation), p-values for the random effects are based on likelihood ratio test. All analysis may be performed on `lmer` object of **lme4** package.

Usage

```
step(model, ddf = "Satterthwaite", type = 3, alpha.random = 0.1, alpha.fixed = 0.05,
reduce.fixed = TRUE, reduce.random = TRUE, fixed.calc = TRUE, lsmeans.calc = TRUE,
difflsmeans.calc = TRUE, test.effs = NULL,  keep.effs = NULL, ...)
```

Arguments

`model`	linear mixed effects model (lmer object).
`ddf`	approximation for denominator degrees of freedom. By default Satterthwaite's approximation. ddf="Kenward-Roger"" calculates Kenward-Roger approximation
`type`	type of hypothesis to be tested (SAS notation). Either type=1 or type=3.
`alpha.random`	significance level for elimination of the random part (for LRT test)
`alpha.fixed`	significance level for elimination of the fixed part (for `F` test and `t`-test for least squares means)
`reduce.fixed`	logical for whether the reduction of the fixed part is required
`reduce.random`	logical for whether the reduction of the random part is required
`fixed.calc`	logical for whether the calculation of the table for fixed effects is needed. If FALSE then only the analysis of random effects is done
`lsmeans.calc`	logical for whether the calculation of LSMEANS(population means) is required
`difflsmeans.calc`	logical for whether the calculation of differences of LSMEANS is required
`test.effs`	charachter vector specifying the names of terms to be tested in LSMEANS. If NULL all the terms are tested. If `lsmeans.calc==FALSE` then LSMEANS are not calculated.
`keep.effs`	charachter vector specifying the names of terms to be kept in the model even if being non-significant
`...`	other potential arguments.

Details

Elimination of all effects is done one at a time. Elimination of the fixed part is done by the principle of marginality that is: the highest order interactions are tested first: if they are significant, the lower order effects are not tested for significance. The step function of lmerTest overrides the one from stats package for lm objects. So if the lmerTest is attached and one wants to call step fof lm object, then needs to use stats::step

Value

`rand.table`	data frame with value of Chi square statistics, p-values for the likelihood ratio test for random effects
`anova.table`	data frame with tests for whether the model fixed terms are significant (Analysis of Variance)
`lsmeans.table`	Least Squares Means data frame with p-values and Confidence intervals
`diffs.lsmeans.table`	Differences of Least Squares Means data frame with p-values and Confidence intervals
`model`	Final model - object of merLmerTest(contains mer class) or gls (after all the required reduction has been performed)

Note

For the random coefficient models: in the random part if correlation is present between slope and intercept, then the simplified model will contain just an intercept. That is if the random part of the initial model is `(1+c|f)`, then this model is compared to `(1|f)` by using LRT. If there are multiple slopes, then the the slope with the highest p-value (and higher then alpha level) is eliminated. That is if the random part of the initial model has the following form `(1+c1+c2|f)`, then two simplified models are constracted and compared to the initial one: the first one has `(1+c1|f)` in the random part and the second one has: `(1+c2|f)`.

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

See Also

`rand`, `lsmeans`, `difflsmeans`

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)

## Not run:
m <- lmer(Informed.liking ~ Product*Information*Gender+
(1|Consumer) + (1|Product:Consumer), data=ham)


#elimination of non-significant effects
s <- step(m)

#plot of post-hoc analysis of the final model
plot(s)
```

```
m <- lmer(Coloursaturation ~ TVset*Picture+
(1|Assessor)+(1|Assessor:TVset), data=TVbo)

step(m, keep.effs = "Assessor")


## End(Not run)
```

---

summary-methods *Methods for Function* summary *in Package* **lmerTest**

Description

Methods for function summary in package **lmerTest**

Methods

signature(object = "merModLmerTest" ,ddf = "Satterthwaite" ,...) summary of the results of linear mixed effects model fitting of object. Returns the same output as summary of "merMod" class but with additional columns with the names "df", "t value" and "Pr(>t)" representing degrees of freedom, t-statistics and p-values respectively calculated based on Satterthwaite's or Kenward-Roger's approximations. summary

Examples

```
(fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy))


## will give you an additional column with p values for the t test
summary(fm1)

##using Kenward-Roger approximations to degrees of freedom
if(require(pbkrtest))
summary(fm1, ddf="Kenward-Roger")

#will give the summary of lme4 package
summary(fm1, ddf="lme4")
```

TVbo *TV dataset*

Description

The TVbo dataset comes from Bang and Olufsen company. The main purpose was to test products, specified by two attributes Picture and TVset. 15 different response variables (characteristics of the product) were assessed by trained panel list.

Usage

```
TVbo
```

Format

`Assessor` factor: numbering identifying assessors

`TVset` factor: attribute of the product

`Picture` factor: attribute of the product

`15 Characteristics of the product` numeric variables: Coloursaturation, Colourbalance, Noise, Depth, Sharpness, Lightlevel, Contrast, Sharpnessofmovement, Flickeringstationary, Flickeringmovement, Distortion, Dimglasseffect, Cutting, Flossyedges, Elasticeffect

Source

Bang and Olufsen company

Examples

```
#import lme4 package and lmerTest package
library(lmerTest)



## Not run:
m <- lmer(Coloursaturation ~ TVset*Picture+
(1|Assessor)+(1|Assessor:TVset), data=TVbo)

step(m, test.effs="TVset", reduce.fixed=FALSE, reduce.random=TRUE)

## End(Not run)
```

Index

Appendix H

Reference manual for the R package SensMixed

Alexandra Kuznetsova, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen (2013) Reference manual for the R package lmerTest

Package 'SensMixed'

January 8, 2015

Type Package

Title Mixed effects modelling for sensory and consumer data

Version 2.0-6

Date 2013-09-10

Author Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

Maintainer Alexandra Kuznetsova <alku@dtu.dk>

Depends stats, lmerTest

Imports Hmisc, gplots, parallel, plyr, lsmeans, doBy, xtable, reshape2, ggplot2

Description
The package provides with functions that facilitate analysis of Sensory as well as Consumer data

License GPL (>= 2)

R topics documented:

`consmixed` *Automated model selection process for the Consumer data*

Description

Constructs the biggest possible model and reduces it to the best by principle of parcimony. First elimination of random effects is performed following by elimination of fixed effects. The LRT test is used for testing random terms, F-type hypothesis test is used for testing fixed terms. The post-hoc and plots are provided

Usage

```
consmixed(response, Prod_effects, Cons_effects=NULL,
Cons, data, structure = 3, alpha.random = 0.1, alpha.fixed = 0.05, ...)
```

Arguments

response	name of the liking variable in the Consumer data
Prod_effects	vector with names of the variables associated with products
Cons_effects	vector with names of the effects associated with consumers
Cons	name of the column in the data that represents consumers
data	data frame (data from consumer studies)
structure	one of the values in c(1,2,3). 1:Analysis of main effects, Random consumer effect AND interaction between consumer and the main effects(Automized reduction in random part, NO reduction in fixed part). 2: Main effects AND all 2-factor interactions. Random consumer effect AND interaction between consumer and all fixed effects (both main and interaction ones). (Automized reduction in random part, NO reduction in fixed part). 3: Full factorial model with ALL possible fixed and random effects. (Automized reduction in random part, AND automized reduction in fixed part).
alpha.random	significance level for elimination of the random part (for LRT test)
alpha.fixed	significance level for elimination of the fixed part (for F test)
...	other potential arguments.

Value

rand.table	table with value of Chi square test, p-values e t.c. for the random effects
anova.table	table which tests whether the model fixed terms are significant (Analysis of Variance)
model	Final model - object of class lmer or gls (after all the required reduction has been performed)

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

Examples

```
library(SensMixed)
data(ham)

consmixed(response="Informed.liking",
Prod_effects= c("Product","Information"),
Cons_effects=c("Gender","Age"), Cons = "Consumer", data =ham, structure=1)
```

convertToFactors *converts variables of the data frame to factors*

Description

the user specifies which variables he/she would like to consider as factors, the functions converts them to factors

Usage

```
convertToFactors(data, facs)
```

Arguments

data	data frame
facs	vector with names of variables that the user would like to convert to factors

Value

returns the same data frame as in the input but with the specified variables converted to factors

Author(s)

Alexandra Kuznetsova

Examples

```
library(SensMixed)
data(carrots)
str(carrots)

carrots <- convertToFactors(carrots, c("Consumer", "Income", "Homesize"))

str(carrots)
```

plot *function creates plots for the sensmixed object*

Description

function creates barplots for the square roots of F statistics and square roots of chi square values for all attributes

Usage

```
## S3 method for class sensmixed
 plot(x, mult = FALSE, dprime = FALSE, sep = FALSE,
                              cex = 2,  interact.symbol = ":", isFixed = TRUE,
                              isRand = TRUE, isScaling = TRUE, ...)
```

Arguments

x	object of class sensmixed
mult	logical. Should multiple plots be plotted, that is barplots for each effect in a separate plot
dprime	logical. Should multiattribute plot for product effects use average squared dprimes instead of square root of F statistics
sep	logical. If TRUE then separate plot is plotted for each effect (mult argument should be then also TRUE)
cex	The magnification to be used
interact.symbol	The symbol to be used for the interaction effects
isFixed	logical. Whether to plot tests of the fixed effects
isRand	logical. Whether to plot tests of the random effects
isScaling	logical. Whether to plot the scaling factor if present
...	other potential arguments.

Value

NULL is returned

Author(s)

Alexandra Kuznetsova

Examples

```
res <- sensmixed(c("Coloursaturation", "Colourbalance"),
                 Prod_effects=c("TVset"),
                 individual="Assessor", data=TVbo, MAM=TRUE,
                 reduce.random=FALSE)
plot(res)
plot(res, mult = TRUE)
plot(res, interact.symbol = " x ")
```

---

plot.consmixed *plots the post-hoc for the consmixed object*

Description

plots the least squares means and differences of least squares means together with the confidence intervals for the fixed effects

Usage

```
## S3 method for class consmixed
 plot(x, main = NULL, cex = 1.4,
                           which.plot = c("LSMEANS", "DIFF of LSMEANS"),
                           effs = NULL, ...)
```

Arguments

x	object of class consmixed
main	string. Title for the plots
cex	A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default
which.plot	type of plot to be drawn
effs	name of the effext for which to draw the plots
...	other potential arguments.

Value

returns NULL

Author(s)

Alexandra Kuznetsova

Examples

```
res <- consmixed(response="Informed.liking",
Prod_effects= c("Product","Information"),
Cons_effects=c("Gender","Age"), Cons = "Consumer", data =ham, structure=1)

plot(res)
```

---

saveToDoc *save the result in tables into a doc file for sensmixed or consmixed objects*

Description

save the tests for the random and fixed effects into a doc file for sensmixed or consmixed objects

Usage

```
saveToDoc(x, file = NA, bold = FALSE, append = TRUE)
```

Arguments

x	object of class sensmixed or consmixed.
file	a character string naming the file to write to, or NULL to stop sink-ing.
bold	logical. Should the significance be in bold text instead of the stars. The default is FALSE
append	logical. If TRUE, output will be appended to file; otherwise, it will overwrite the contents of file.

Author(s)

Alexandra Kuznetsova

Examples

```
## Not run:
res <- sensmixed(c("Coloursaturation", "Colourbalance"),
                 Prod_effects=c("TVset"),
                 individual="Assessor", data=TVbo)

saveToDoc(res, file = "C:/Desktop/output.doc")

## End(Not run)
```

`sensmixed` *Automated model selection process for each attribute of sensory data*

Description

Constructs the biggest possible models for the selected attributes and reduces them to the best by principle of parsimony models. First elimination of random effects is performed following by elimination of fixed effects. The LRT test is used for testing random terms, F-type hypothesis test is used for testing fixed terms

Usage

```
sensmixed(attributes=NULL, Prod_effects, replication = NULL,
                              individual, data, product_structure = 3,
                              error_structure ="No_Rep", MAM = FALSE,
                              mult.scaling = FALSE,
                              MAM_PER = FALSE, adjustedMAM = FALSE,
                              alpha_conditionalMAM = 1,
                              calc_post_hoc = FALSE, parallel = FALSE,
                              reduce.random=TRUE, alpha.random = 0.1,
                              alpha.fixed = 0.05, interact.symbol = ":", ...)
```

Arguments

`attributes`	vector with names of sensory attributes
`Prod_effects`	names of the variables related to the product
`replication`	names of the replication column in the data, if present
`individual`	name of the column in the data that represent assessors
`data`	data frame (data from sensory studies)
`product_structure`	one of the values in c(1, 2, 3). 1: only main effects will enter the initial biggest model. 2: main effects and 2-way interaction. 3: all main effects and all possible interaction
`error_structure`	one of the values in c("No_Rep", "2-WAY", "3-WAY"). "No_Rep" and "2-WAY" - assessor effect and all possible interactions between assessor and Product_effects. "3-WAY" - assessor and replicate effect and interaction between them and interaction between them and Product_effects
`MAM`	logical. if MAM model should be constructed (scaling correction)

mult.scaling	logical. Whether multiple scaling should be used
MAM_PER	logical. if MAManalysis function should be called (scaling correction)
adjustedMAM	logical. should MAM be adjusted for the scaling
alpha_conditionalMAM	logical. scaling should be part of the model in case its p-value is less than alpha_conditionalMAM
calc_post_hoc	logical. Should the post hoc analysis be performed on the final resuced models for all the attributes
parallel	logical. Should the computation be done in parallel. the default is TRUE
reduce.random	logical. Eliminate non-significant random effects according to alpha.random or not. The default is TRUE
alpha.random	significance level for elimination of the random part (for LRT test)
alpha.fixed	significance level for elimination of the fixed part (for F test)
interact.symbol	symbol for the indication of the interaction between effects. the default one is ":".
...	other potential arguments.

Value

FCHi	matrix with Chi square values from LRT test and F values form F-type test for the selected attributes
pvalue	matrix withp-values for all effects for the selected attributes

Author(s)

Alexandra Kuznetsova, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen

Examples

```
#import SensMixed package
library(SensMixed)

#import TVbo data from lmerTest package
data(TVbo)

#run automated selection process
res <- sensmixed(c("Coloursaturation", "Colourbalance"),
Prod_effects = c("TVset", "Picture"), replication="Repeat",
individual="Assessor", data=TVbo, MAM=TRUE)

res_paral <- sensmixed(names(TVbo)[5:ncol(TVbo)],
Prod_effects = c("TVset", "Picture"), replication="Repeat",
individual="Assessor", data=TVbo, error_structure="3-WAY")


## run MAManalysis function
res_MAM <- sensmixed(c("Coloursaturation", "Colourbalance"),
                Prod_effects=c("TVset"), replication="Repeat",
                 individual="Assessor", data=TVbo, MAM_PER=TRUE)
```

```
## print is not yet implemented
## get anova part
res_MAM[[3]][,,1]

## compare with the general implementation
res <- sensmixed(c("Coloursaturation", "Colourbalance"),
                 Prod_effects=c("TVset"),
                 individual="Assessor", data=TVbo, MAM=TRUE,
                 reduce.random=FALSE)
res$fixed

## Not run:
plot F and Chi square values
plot(result)

## End(Not run)


result <- sensmixed(names(TVbo)[5:ncol(TVbo)],
Prod_effects=c("TVset", "Picture"),
replication="Repeat", individual="Assessor", data=TVbo,
calc_post_hoc = TRUE)

result
result$fixed

result_MAM <- sensmixed(names(TVbo)[5:ncol(TVbo)],
Prod_effects=c("TVset", "Picture"),
replication="Repeat", individual="Assessor", data=TVbo,
MAM = TRUE)

result_MAM

result_MAM_mult <- sensmixed(names(TVbo)[5:ncol(TVbo)],
Prod_effects=c("TVset", "Picture"),
replication="Repeat", individual="Assessor", data=TVbo,
MAM = TRUE, mult.scaling = TRUE)

result_MAM_mult
```

Index

Appendix I

Tutorial for the SensMixed package

Kuznetsova A. Brockhoff P.B.B.

Tutorial for the SensMixed application

Alexandra Kuznetsova, Per Bruun Brockhoff

1. The SensMixed package - an overview

The SensMixed package is an **R** package for analysing Sensory and Consumer data in a mixed model framework developed by Alexandra Kuznetsova, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen. The package facilitates, among other things:

- analysis of sensory data in a mixed model framework
- novel tools for correcting scaling effects in sensory data
- analysis of consumer data in a mixed model framework
- post-hoc analysis
- plots
- ready to publish output

The package also provides a graphical user interface (GUI), which is based on shiny **R** package Chang et al. (2015), so that it becomes very user friendly and easy to use for sensory practitioners.

2. Obtaining SensMixed

You have to have the latest **R** program installed (at least 3.2-0) on your computer before you think about the SensMixed package. You can download and install the latest version of **R** from `http://cran.r-project.org/`. Once you have installed **R** you start using it, but many people (and I encourage that) want to install a suitable GUI (graphical user interface) or IDE (integrated development en- vironment). One suitable choice is the (also free) program RStudio, which you can download and install from `http://www.rstudio.com/ide/download/desktop`.

Finally you can install the SensMixed package by clicking in the Install button and writing SensMixed in Rstudio or write in the **R** console the following command:

```
install.packages("SensMixed")
```

In order to use the functions from the package you need to attach the package by writing the following in your **R** console:

```
library("SensMixed")
```

If you encounter some installation problems, then you are more than welcome to contact the maintainer of the package Alexandra Kuznetsova (alku@dtu.dk).

3. The graphical user interface

The SensMixed package contains a shiny application Chang et al. (2015), that provides a graphical user interface for the functions contained in the SensMixed. In order to launch the application, one simply needs to run the following line in the **R** console:

```
SensMixedUI()
```

This command launches the application in your default web browser. The application supporting the package was designed with focus on simplicity and usability such that valuable information may be accessed in an easy way. Figure 1 represents the main widget of the GUI. In the *Choose type of analysis* panel to the left you can specify which type of analysis you would like to perform. There are two options: analysis of sensory or consumer data. The left panel consists of three tabs: *Input arguments*, *Modeling controls* and *Analysis controls*. In the *Input arguments* you select the names of the variables, that you would like to analyze - these variables are coming from the data. *Modeling controls* tab stands for a detailed specification of the type of modelling. The *Analysis controls* tab stands for specification of the type of analysis to be performed. These tabs are described in details in the following sections. The tabs at the top right of the widget are: *Data*, *Plot output*, *Table output*, *Step output*, *Post-hoc* and *MAM analysis*. The first one stands for the import of the data and is the one that is selected in the figure, the other five tabs are dedicated for the output of results from the analysis of sensory data. In the following sub sections each tab will be explained in detail. First, however, I will explain how the data is imported into the application.

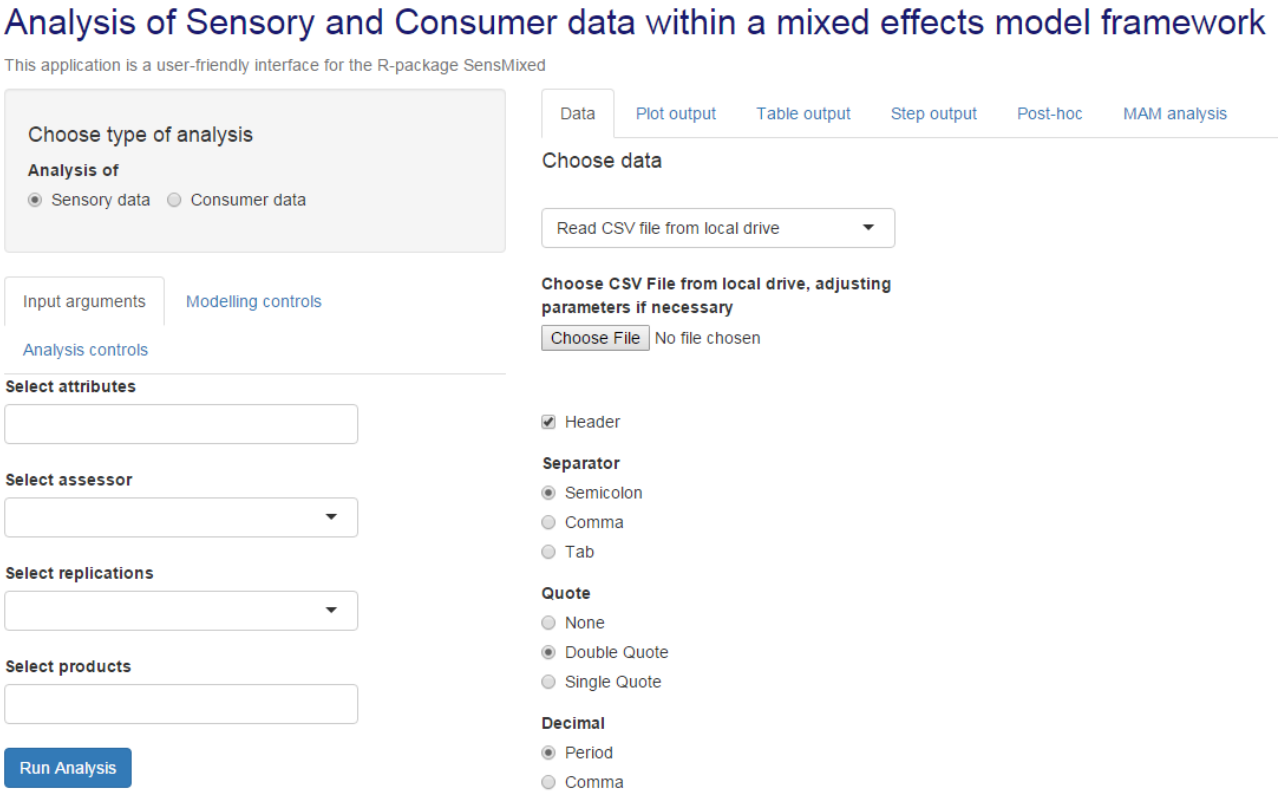


Figure 1: The main widget of the GUI of the SensMixed application

4. Data import

In Figure 1 the screenshot for the data import is displayed. As can be seen the user may choose, which data to use: either to import the data from the local files, or the user may choose two data sets, that are contained in the SensMixed: the `TVbo` data coming from the sensory studies and the `ham` data coming from the consumer studies - these two data sets I am going to use in this tutorial. The user is always welcome to play around with them.

Even if the first option for choosing the data says: *Read CSV file from local drive*, it accepts different formats:

- plain files such as .txt, .csv
- Excel files such as .xls, xlsx

The type of format is chosen through the *Separator* box. The details of import can be controlled: for example, whether to include a header in the imported data, or which type of decimal to use and others. These options give the flexibilities to import different data.

For illustrative purposes I will use here the `TVbo` data, which is a sensory data, contained in the SensMixed package. The `TVbo` data was produced by the highend HIFI company Bang and Olufsen A/S, Struer, Denmark, and was used for a workshop at the 8th Sensometrics Meeting in Norway in 2008. In this data the main purpose was to assess 12 products, specified by two

features: Picture (factor with 4 levels) and TVset (factor with 3 levels). The products were assessed by 8 assessors in 2 replications for 15 different attributes.

The first and foremost step in every analysis is to get the data that is to be analyzed.

5. Choose data

In the *Choose data* tab I select `TVbo`. Figure 9 shows the screenshot of the chosen data (here `TVbo`).

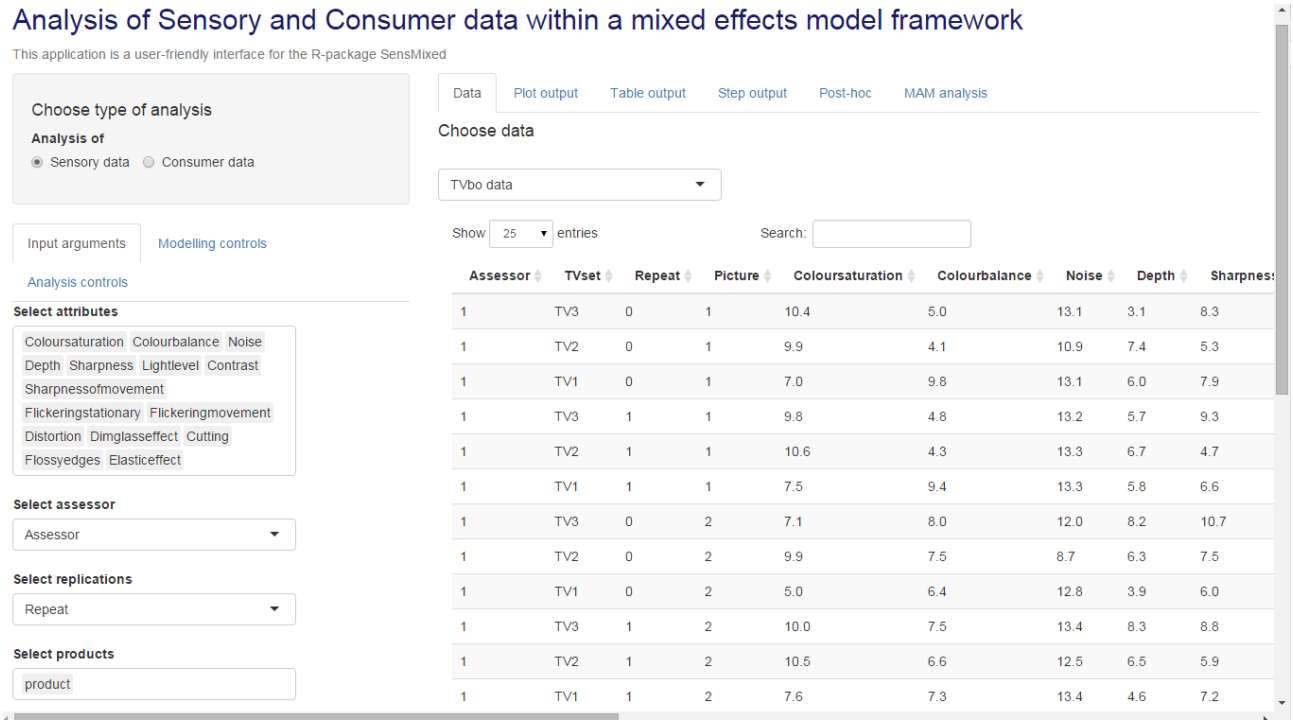


Figure 2: Screenshot for the TVbo data and input arguments for the analysis of the TVbo data

From Figure 9 it can be seen that the `TVbo` data contains column with the name `Assessor`, which has numbers for each of the eight Assessors. `TVset` column stands for the `TVset` feature, which has three levels: `TV1`, `TV2` and `TV3`. `Repeat` column contains two numbers (0 and 1) referring to the number of replication. `Picture` column stands for the `Picture` feature and contains numbers referring to the levels of the Picture feature: 1,2,3 and 4. `product` column referring to the 12 tested products (3 by 4 TVset Picture combinations). The rest of the columns are the assessors' scores for 15 attributes.

6. Input arguments

In the *Input arguments* tab to the left you need to put the names of the variables for the analysis. For the TVbo data the arguments are automatically filled by the program (see Figure 9):

Select attributes: the names of the columns in the data corresponding to the scores for the attributes. `Coloursaturation`, `Colourbalance`, `Noise` and the rest of 15 attributes

Select Assessor: the name of the column corresponding to the assessors in the data (`Assessor`)

Select replications: the name of the column corresponding to the replication in the data (`Repeat`)

Select products: the names of the product factors. (`product`).

Before clicking the *Run Analysis* button, you may specify what type of analysis you would like to perform via the *Modelling controls* tab.

7. Modelling controls

The analysis of the sensory data in SensMixed is performed in a mixed effects model framework. For each attribute a linear mixed effects model is constructed by using the `lmer` function from the lme4 package Bates et al. (2013) for constructing the mixed effects models and Kuznetsova et al. (2013) for testing the effects for significance . The *Modelling controls* tab stands for the specification of the type of model to be considered.

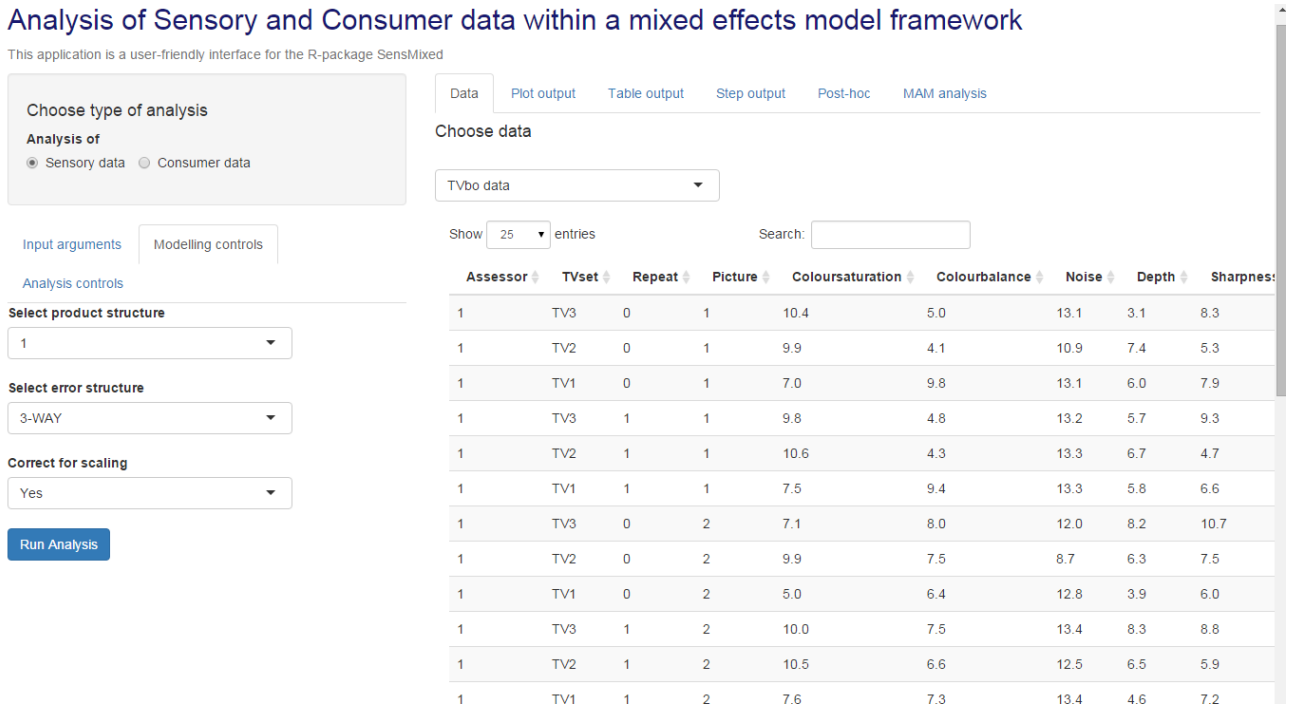


Figure 3: Screenshot for the Modelling controls for the TVbo data

Figure 3 represents the screenshot for the *Modelling controls* tab. The inputs in this screenshot stand for the specification of a linear mixed effect model for each sensory attribute. In the following I will describe each of the input for the specification of linear mixed effects models in general and will explain later on which inputs I have chosen for the analysis of the `TVbo` data.

Select product structure: This tab stands for the specification of the fixed effects in linear mixed effects models. There are three following options:

1 only main effects form the fixed part

2 main effects and 2-way interactions form the fixed part

3 all main effects and all possible interactions form the fixed part

Select error structure: This tab stands for the specification of the random effects in linear mixed effects models. There are also three options here:

No-Rep assessor effect and all possible interactions between assessor and fixed effects

2-WAY No-Rep option plus replicate effect and replicate assessor interaction effect

3-WAY assessor and replicate effect and interaction between them and interaction between them and all fixed effects

Correct for scaling:

Yes consider mixed assessor model Brockhoff et al. (2015), where additional fixed effect is added, standing for the scaling effect

No do not consider mixed assessor model

8. Analysis controls

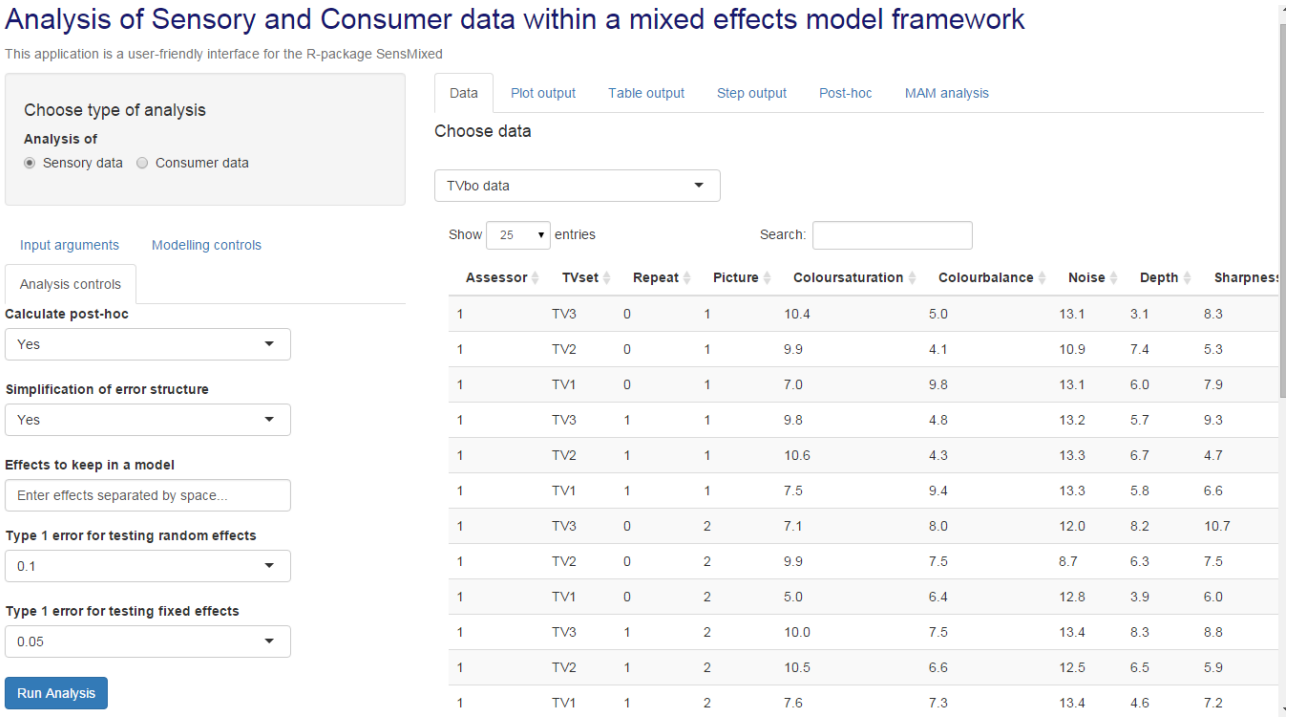


Figure 4: Screenshot for the Analysis controls for the TVbo data

The input tabs in Figure 4 stand for the specification of the type of analysis to be performed on models defined in input tabs in Figure 3 and are the following ones:

Simplification of error structure: This tab stands for the simplification of the random effects in linear mixed effects models. There are also two options here:

Yes sequentially eliminate non-significant random effects following procedure proposed in Kuznetsova et al. (2015) using the Type 1 error rate (0.1 the default one)

No do not eliminate random effects

Effects to keep in the model: Here one needs to type the effects that one would like to keep in the model even if not being significant. By default *Assessor*) and highest order interaction between `Assessor` and product effects (here `product:Assessor`) are always kept in the model.

Type 1 error rate for testing random effects:

0.1 this option is recommended in Kuznetsova et al. (2015)

0.2

0.05

Type 1 error rate for testing fixed effects:

0.05 this option is recommended in Kuznetsova et al. (2015)

0.01

0.001

9. Analysis of TVbo sensory data. No scaling correction

For illustration purposes I have chosen here *product structure = 1*, which considers only one product main effect (`product`). For the error specification part I have chosen *error structure = 3-WAY*, which considers the maximal possible error structure for the initial model, that is all possible random effects and interactions between random and fixed effects. According to the chosen controls for the model specification, the following linear mixed model is constructed for each attribute:

$$y_{ijk} = \mu + a_i + \nu_j + d_{ij} + r_k + ar_{ik} + a\nu_{jk} + \varepsilon_{ijk} \tag{1}$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{assessor \times product}),\ r_k \sim N(0, \sigma^2_{replicate}),$$
$$ar_{ik} \sim N(0, \sigma^2_{assessor \times replicate}),\ \nu r_{jk} \sim N(0, \sigma^2_{product \times replicate}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

Then the inputs standing analysis controls needs to be chosen. Here, as can be observed from Figure 4, in the *Simplification of error structure* input I have put the default one *Yes*, which eliminates sequentially non-significant random effects as suggested by Kuznetsova et al. (2015). Note that the random effects `Assessor` and `Assessor` × `Product` are always kept in the model. Finally, I choose the default numbers for the Type 1 error rates.

In order to view the results, I click on the *Plot output* tab (similarly one may open any of the results tabs, but only the first time the analysis is run. Then one needs to click on the *Run analysis* button).

9.1. Plot output

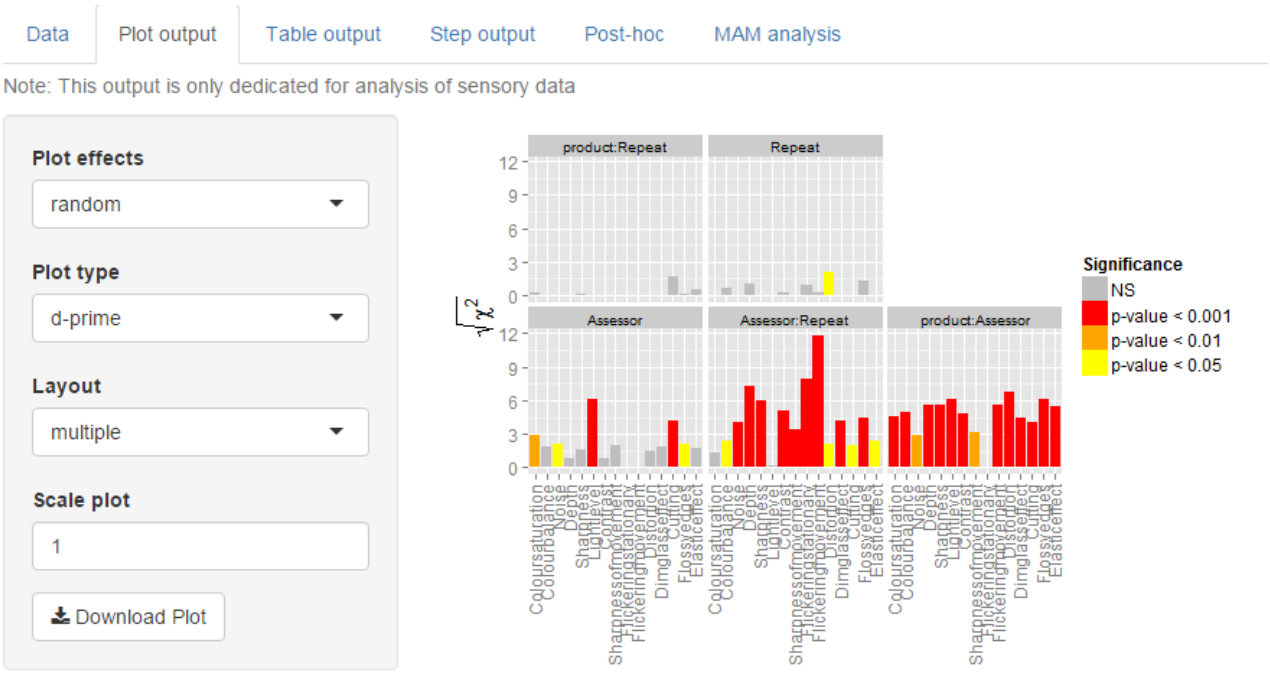


Figure 5: Screenshot for the plot output for random effects for the TVbo data

Figure 5 shows the screenshot of the *Plot output* for analysis of random effects for the `TVbo` data. In the panel to the left I have chosen the plot for the random effects. Layout multiple means that I get multiple subplots for each random effect, The Plot shows the bars for the sequential χ^2 statistics of the likelihood ratio test applied to each random effect for each attribute. The sequential means that the χ^2 values come from the stepwise selection process of elimination of non-significant random effects. The x-axis stands for the attributes. An overview of the plot indicates that there is no replicate effect in the data. It seems like there is a disagreement between assessors in scoring the products (`product:Assessor` effect is significant for almost all the attributes). With the *Download plot* one may easily save the plot in the .pdf format to the local disc. With the *Scale plot* input one may scale the plot - this is valuable when downloading the plot: if it becomes too big, then one may write *2* in the *Scale plot* input and then try to download the plot again.

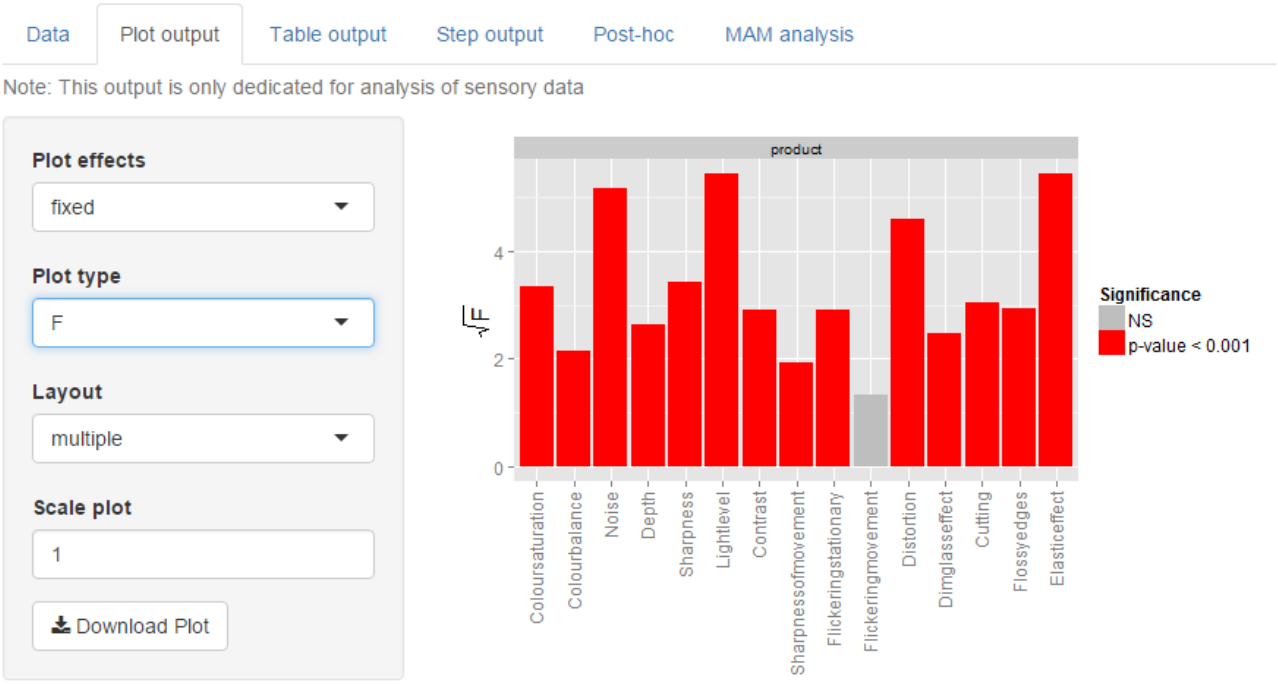


Figure 6: Screenshot for the plot output for the fixed effects for the TVbo data

Finally one turns to the results from the analysis of the fixed effects. Figure 6 shows the screenshot for the *Plot output* of the analysis of the fixed effects (`product` effect). From the plot it can be observed, that the `product` effect is significant for all attributes except `Flickeringmovement`.

9.2. Table output

The *Table output* tab, which is next to the *Plot output* tab provides the same output as *Plot output* just in tables.

9.3. Step output

The *Step output* tab provides a detailed information on the analysis of the fixed and random effects. Figure 7 shows the screenshot for the *Step output* tab for the analysis of the `Colourbalance` attribute from the `Tvbo` data. In the panel to the left there is a selection list input: here one may choose for which attribute one wants to view the results of the analysis. Here I have chosen attribute `Colourbalance`. The first table at the top presents the analysis of the random effects. In the `elim num` column one may view the order in which non-significant effects were eliminated as being non-significant according to the chosen Type 1 error rate (here 0.1). `kept` means the effect was kept in the model. It can be seen that 2 random effects were eliminated: `product:Repeat` and `Repeat`. The next table represents the analysis of the fixed effects. It can be seen that the `product` effect is highly significant, which means that the assessors can discriminate the products according to `Colourbalance` attribute. One can easily save

the table to the local disc via *Download Table* button. One may choose to download in *html* format or in *latex* format.

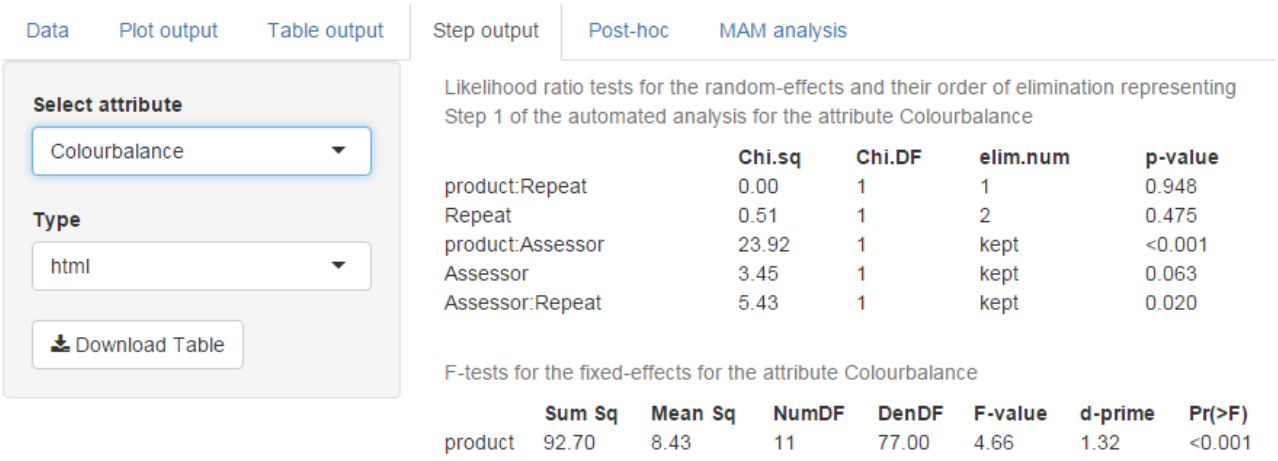


Figure 7: Screenshot for the step output for the Colourbalance attribute for the TVbo data

9.4. Post-hoc output

The *Post hoc* tab represents the results of the post-hoc analysis, namely pairwise-comparison tests for the fixed effects. The Figure 8 represents the screenshot for the *Post hoc* tab for the the attribute `Colourbalance`. In the panel to the left there is a selection list input: here one may choose for which attribute one wants to view the results of the analysis. Here I have chosen the attribute `Colourbalance`. The selection input list *Type of Plot* has a few options: *DIFF of LSMEANS* showing the differences of least squares means for an effect in question and *LSMEANS* showing the least squares means of an effect in question. In selection list *Effects* one can select for which effect to view the results (in this example there is only one `product` effect). The results are displayed in barplots and table, both can be easily downloaded via *Download Table* and *Download Plot* buttons to the local disc. In this example the results are quite hard to interpret since there are 12 products (so $12*11/2 = 66$ comparisons are visualized). Considering multi-way product structure, that is instead of only one `product` effect consider two main effects `TVset` and `Picture` and interaction between them `TVset:Picture`, can simplify interpretation and get more insight into the data (Kuznetsova et al., 2015).

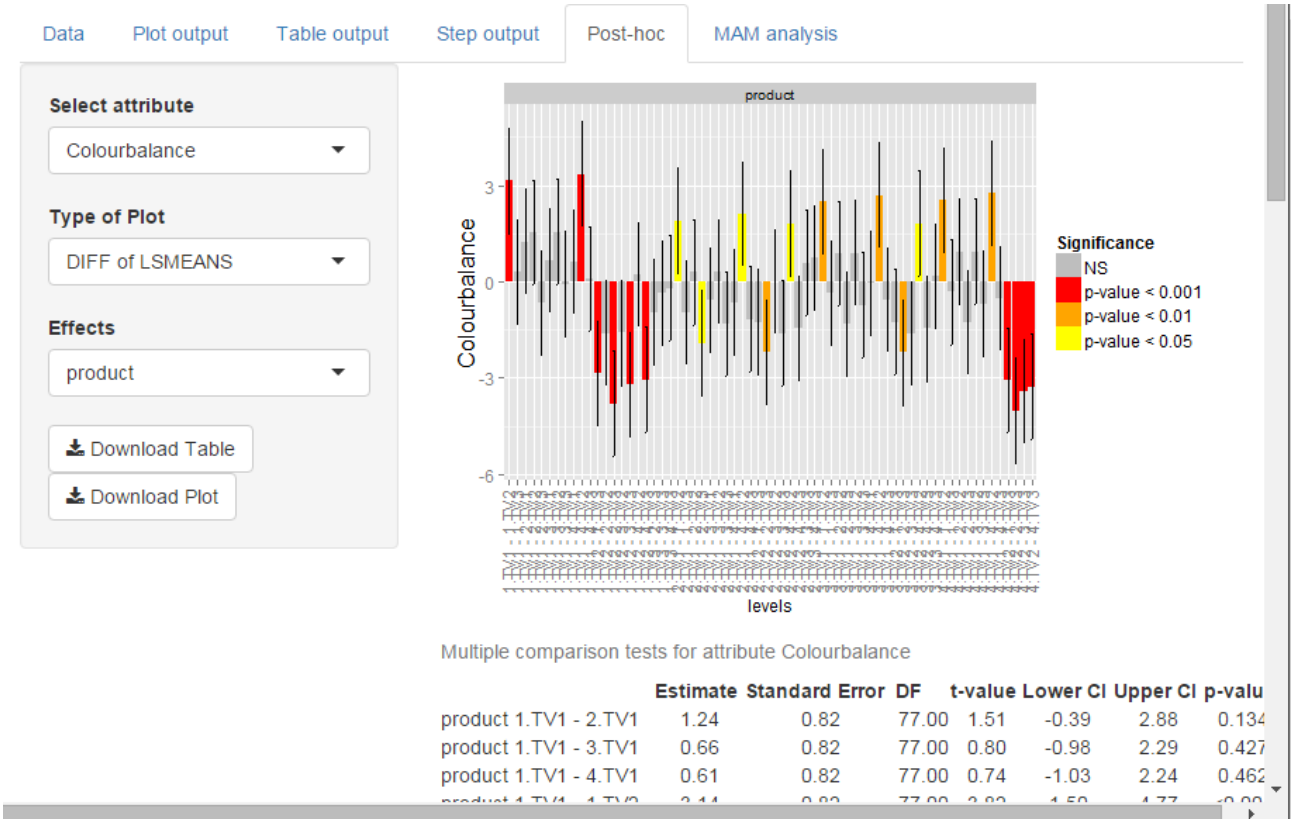


Figure 8: Screenshot for the post-hoc output for the Colourbalance attribute for the TVbo data

10. Analysis of sensory data including scaling correction

In this example the focus is on illustrating how to correct for the scaling effect in the sensory data and how to make the d-tilde plots using the **SensMixed**. For illustrative purposes I will again use here the `TVbo` data.

In the *Choose data* tab I again select `TVbo` (see Figure 9).

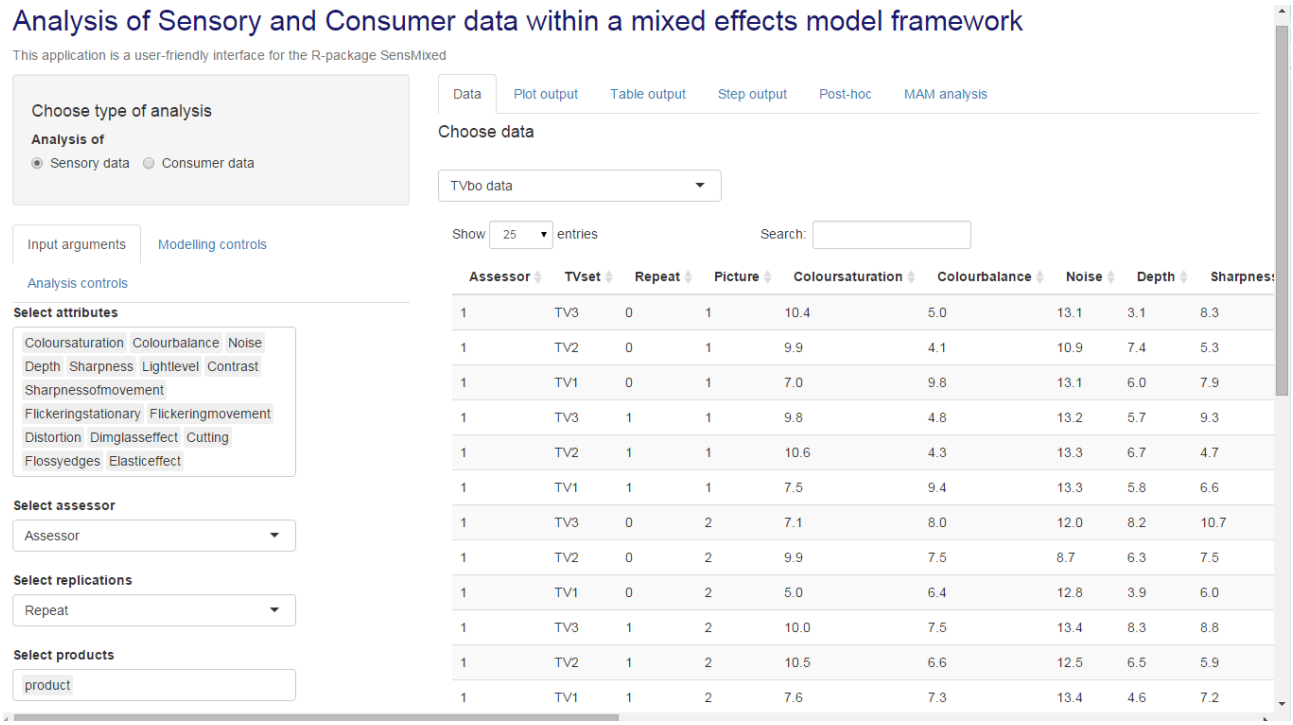


Figure 9: Screenshot for the TVbo data and input arguments for the analysis of the TVbo data

The *Input arguments* are then automatically filled, where as product effects one-way factor `product` is chosen. In the *Modelling controls* tab (see Section 7) I select *product structure = 1*, which considers only one product main effect (`product`). Note, that the *MAM analysis* tab view results for only one product effect with at least 3 levels and only for the balanced data. For the error specification part I have chosen *error structure = No-Rep*, which considers assessor random effect and interaction between assessor and product effect. In order to correct for the possible scaling effects, I select *Correct for scaling = Yes*. For more details about *Input arguments*, *Modelling controls* and *Analysis controls* tabs see Sections 6 , 7, 8. According to the chosen controls for the model specification, the following mixed assessor model (MAM) is constructed for each attribute:

$$Y_{ijk} = \mu + a_i + \nu_j + \underbrace{\beta_i x_j}_{\text{scaling}} + \underbrace{d_{ij}}_{\text{disagreement}} + \varepsilon_{ijk} \tag{2}$$

$$a_i \sim N(0, \sigma^2_{assessor}),\ d_{ij} \sim N(0, \sigma^2_{disagreement}),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

where a_i is the assessor main effect, $i = 1, 2, ..., I$, the ν_j the product main effect, $j = 1, 2, ..., J$, $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and hence β_i is the

individual (scaling) slope (the restriction $\sum_{i=1}^{I} \beta_i = 0$ is imposed in order to ensure that model 2 is uniquely parametrized). The d_{ij} term here captures interactions that are not scale differences hence "disagreements". In (Brockhoff et al., 2015) it was shown that MAM produces valid and improved hypothesis tests for as well overall product differences as post-hoc product difference testing.

Then the inputs standing for analysis controls need to be chosen. Here I have put the default one *Simplification of error structure = Yes* , which eliminates sequentially non-significant random effects. However, the random effects `Assessor` and `Assessor` × `Product` are always kept in the model, so in this example there will be no elimination of random effects done. Finally, I choose the default numbers for the Type 1 error rates.

In order to view the results, I click on the *Plot output* tab (similarly one may open any of the results tabs, but only the first time the analysis is run. After the second time, one needs to click on the *Run analysis* button whenever the analysis needs to be rerun for the selected inputs).

10.1. Plot output

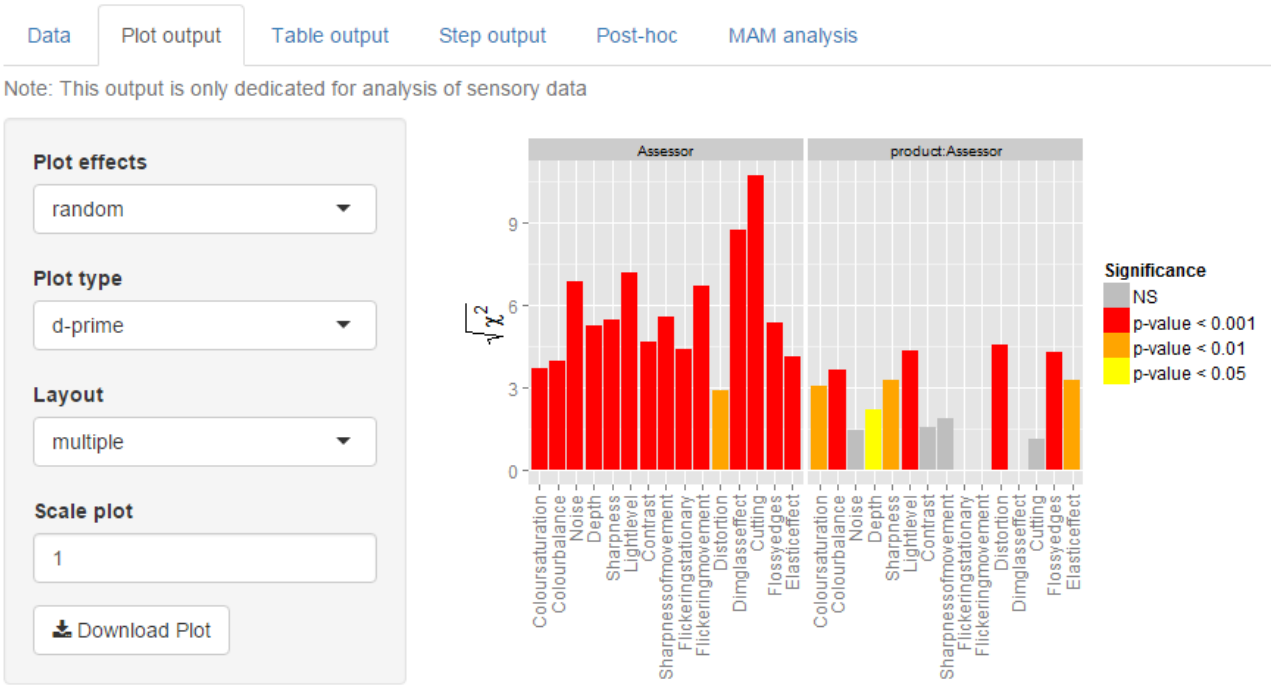


Figure 10: Screenshot for the plot output for random effects for the TVbo data. Scaling corrected

Figure 10 shows the screenshot of the *Plot output* for analysis of random effects for the `TVbo` data. In the panel to the left I have chosen the plot for the random effects. *Layout = multiple* means that I get multiple subplots for each random effect. The plot shows the bars for the χ^2 statistics of the likelihood ratio test applied to each random effect for each attribute. The x-axis stands for the attributes. An overview of the plot indicates that there is a significant `Assessor` effect for all attributes. Not for all attributes there is a significant `Assessor:Product`

interaction. If we compare with the plot in Figure 5, we may notice that the `product:Assessor` effect has become either lower or non-significant for the attributes. This is actually because here the scaling effect is accounted in the models (some of the `product:Assessor` interaction is now taken by the scaling effect, so that `product:Assessor` now represents the pure disagreement between assessors in scoring the products). With the *Download plot* one may easily save the plot in the .png format to the local disc. With the *Scale plot* input one may scale the plot - this is valuable when downloading the plot: if it becomes too big, then one may write *2* in the *Scale plot* input and then try to download the plot again.

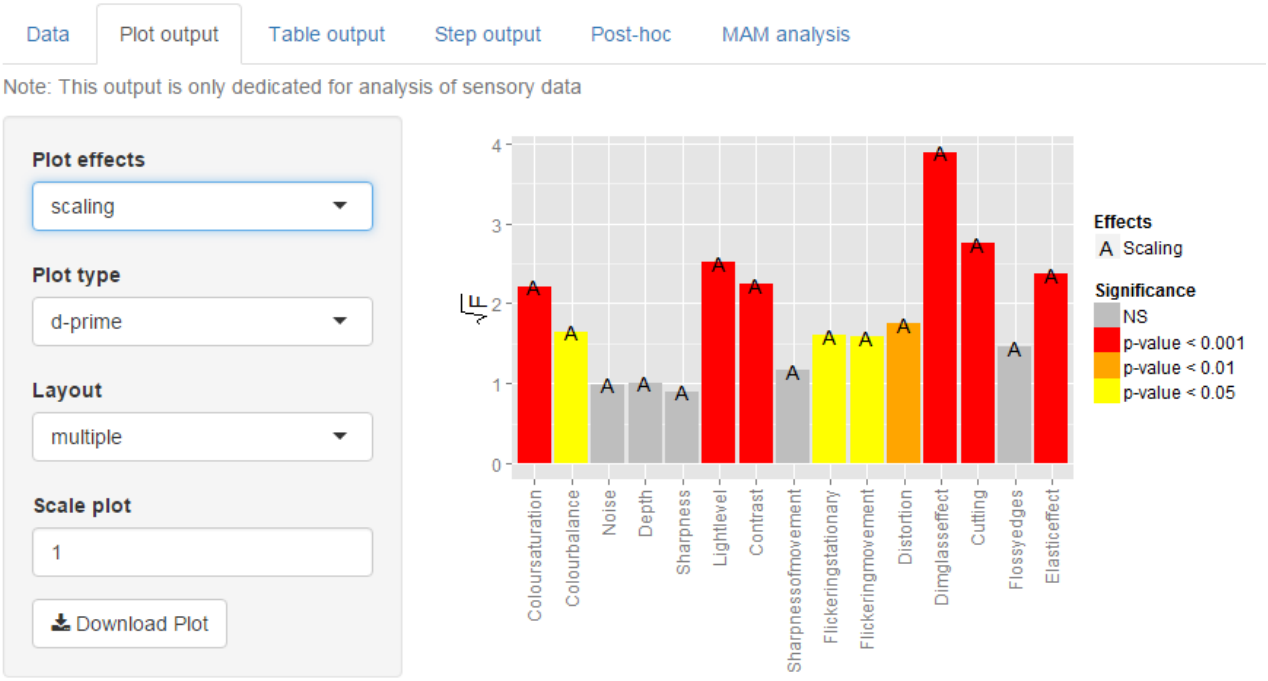


Figure 11: Screenshot for the plot output for the scaling effects for the TVbo data

Next, I take a look at the `Scaling Plot`. The plot shows the bars for the $\sqrt{F}$ statistics of the F test applied to scaling effect for each attribute. The x-axis stands for the attributes. Figure 11 shows the screenshot for the *Plot output* of the analysis of the scaling effects. From the plot it seems like for the majority of attributes the scaling effect is present, hence should be accounted for.

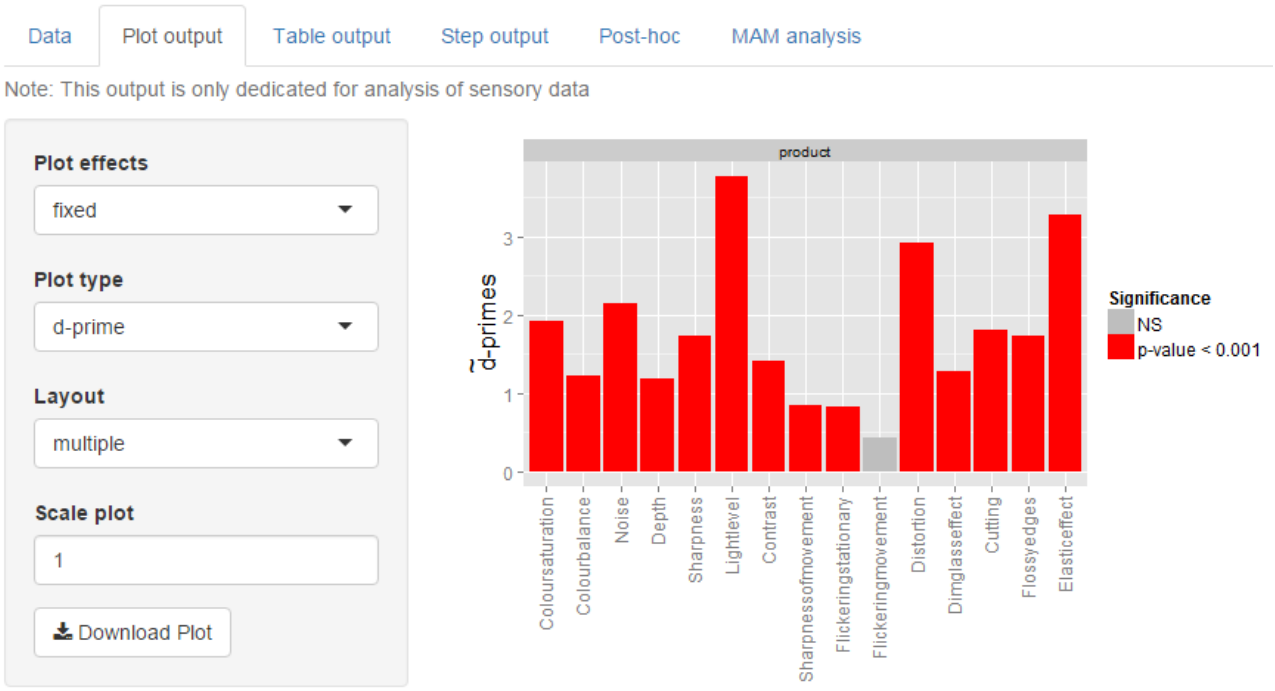


Figure 12: Screenshot for the plot output for the fixed effects for the TVbo data

Finally, I take a look at the analysis of the fixed effect. The plot shows the bars for the $\hat{\tilde{d}}$ of the fixed effect (here `product`) for each attribute. The x-axis stands for the attributes. Figure 12 visualizes the screenshot of the $\hat{\tilde{d}}$ in **SensMixed**. From the plot it can be observed, that the `product` effect is significant for all attributes except `Flickeringmovement`. Since $\tilde{d}$ represents the effect sizes, the sizes of the bars can be directly compared between the attributes. For example, the size of the `product` effect is the highest for `Lightlevel` attribute. This plot is especially valuable for the multi-way product structure situations.

10.2. Step output

The *Step output* tab provides a detailed information on the analysis of the fixed and random effects. Figure 13 shows the screenshot for the *Step output* tab for the analysis of the `Colourbalance` attribute from the `TVbo` data. In the panel to the left there is a selection list input: here one may choose for which attribute one wants to view the results of the analysis. Here I have chosen attribute `Colourbalance`. The first table at the top presents the analysis of the random effects. In the `elim num` column one may view the order in which non-significant effects were eliminated as being non-significant according to the chosen Type 1 error rate (here 0.1). `0` means the effect was kept in the model. It can be seen that all the random effects are kept in the model. The next table represents the analysis of the fixed effects. It can be seen that both `product` and `Scaling` effects are highly significant. The p values, actually, became even lower compared to the model without the `Scaling` effect. One can easily save the table to the local disc via *Download Table* button. One may choose to download in *html* format or in *latex* format.

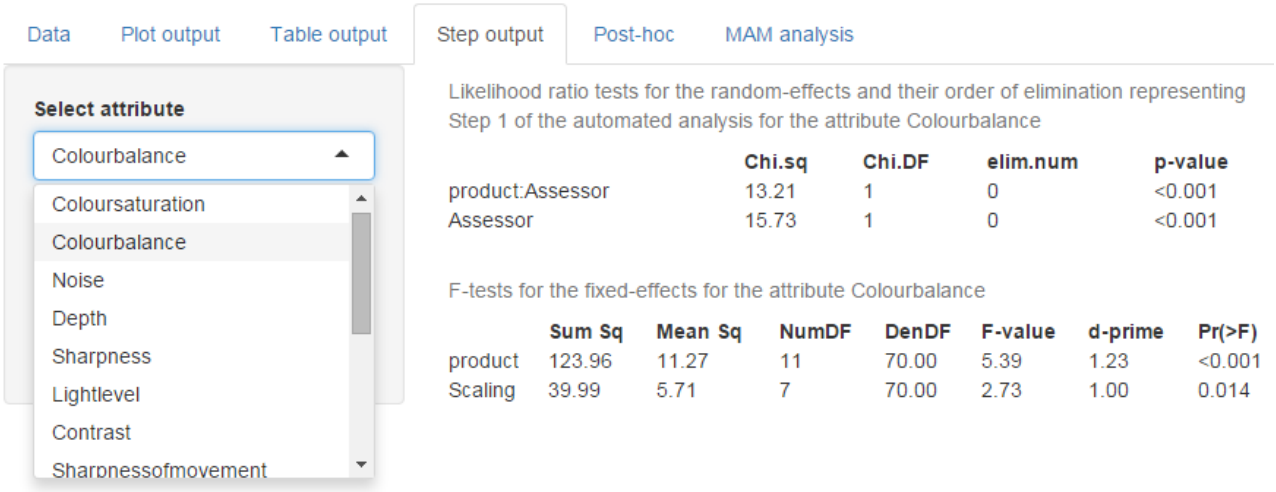


Figure 13: Screenshot for the step output for the Colourbalance attribute for the TVbo data

10.3. Post-hoc output

The *Post hoc* tab represents the results of the post-hoc analysis, namely pairwise-comparison tests for the fixed effects. The output is similar to the one presented in Figure 8 with no correction of the scaling effect. he difference is only in the standard errors and p values in pairwise comparisons. Whenever the scaling effect is significant, the standard errors become lower as well as p values, so the tests become more powerful (as also emphasized in Brockhoff et al. (2015))

10.4. MAM analysis output

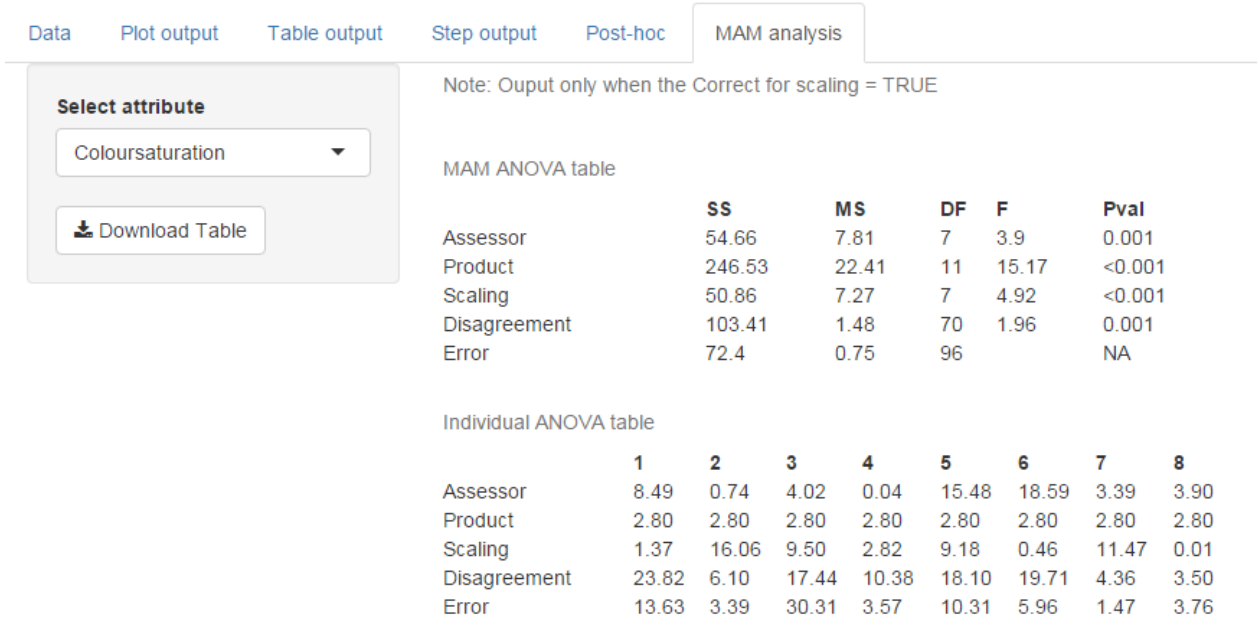


Figure 14: Screenshot for the MAM analysis GUI

In Figure 14 the GUI for the *MAM analysis* tab is visualized. It can be seen that the output is presented for the selected attribute (here `Coloursaturation`). *MAM analysis* tab has some important limitations:

- can only handle balanced data
- can only consider one product factor with at least 3 levels
- can not consider complex error structures (session / batch / carry-over effects e.t.c.)

If one of these requirements is not fulfilled, the the *MAM analysis* tab simply will not produce results. There are a number of table outputs produced for each attribute. In the following these tables will be discussed in details.

10.4.1. MAM ANOVA table

This output is almost the same as the one coming from the *Step output* and gives the overall ANOVA table. One may even check that the F and p values for the scaling and product effects are identical. From Figure 14 it can be observed that both product and scaling effects are significant for the `Coloursaturation` attribute.

10.4.2. Individual ANOVA table

individually decomposed ANOVA table for each attribute, cf. Table 2 in Brockhoff et al. (2015).

10.4.3. Individual performance tests

Figure 15 presents the creenshot for the individual performance tests for the attribute `Coloursaturation` corresponding to Table 2 of Peltier et al. (2014).

Individual performance tests

	1	2	3	4	5	6	7	8	AVE
Product	1.86 *	1.65 *	1.82 *	4.38 ***	3.12 ***	3.27 ***	2.59 **	3.21 ***	2.74
Scaling	0.79	0.28 ***	1.56 *	1.3	1.55 *	1.12	0.39 ***	1.02	1
Disagreement	1.54	0.78	1.32	1.02 *	1.35	1.4 *	0.66 *	0.59	1.08
Repeatability	1.07	0.53	1.59 ***	0.55	0.93	0.71	0.35	0.56	0.78
Level	0.59 *	-0.18	-0.41	0.04	0.8 **	-0.88 **	-0.38	0.4	0
Correlation	0.67 *	0.53 .	0.9 ***	0.91 ***	0.9 ***	0.81 **	0.72 **	0.95 ***	0.8

Figure 15: Screenshot for the MAM preference tests for Coloursaturation attribute

P-values are categorized using the usual R-symbols:

- " " = p-value ≥ 0.1
- "." = p-value < 0.1
- "*"= p-value < 0.05
- "**" = p-value < 0.01
- "***" = p-value < 0.001

The first four (double) row of these result matrices show the results corresponding exactly to the four rows of Table 2 in Peltier et al. (2014). The MAM-CAP table as such is NOT produced. Instead a descriptive statistic is given for each performance measure. In the following the explanation for each row of the table is provided:

- Product: The square root of the individual assessor product F
- Scaling: The individual beta values (averaging to 1)
- Disagreement: The individual disagreement statistic.
- Repeatability: The individual error (within product) standard deviation
- In addition we provide two more statistics and hypothesis tests:
  - Level: The main effect of assessor (summing to zero) We test whether the individual is different from the average

– Correlation: The correlation between the individual product averages and the overall(consensus) product average. These will average to something close to the so-called Cronbach's Alpha. We test whether the correlation is different from zero, i.e. it can also be seen as the significance test for negativity (if the correlation is in fact negative)

For more details about the Individual performance tests see Peltier et al. (2014) and the Appendix in Section 13. From Figure 15 it can be seen that the panel is discriminative for the attribute `Coloursaturation` (all the p values for the `Product` effect are less than 0.05). Assessors 2 and 7 seem to use lower scale whereas Assessors 3 and 5 use the upper scale. Regarding the `Disagreement`, assessors 4 5 and 7 seem to disagree with the rest of the panel for the attribute `Coloursaturation`.

10.4.4. MAM based post hoc

Pairwise product differences

	Estimate	Standard Error	Pval	Lower CI	Upper CI
1.TV1 - 2.TV1	-0.169	0.43	0.696	-1.0849	0.7138
1.TV1 - 3.TV1	-0.312	0.43	0.470	-1.2477	0.5603
1.TV1 - 4.TV1	-1.044	0.43	0.018	-2.1203	-0.1849
1.TV1 - 1.TV2	-2.681	0.43	<0.001	-4.2288	-1.6521
1.TV1 - 2.TV2	-2.525	0.43	<0.001	-4.026	-1.5244
1.TV1 - 3.TV2	-2.006	0.43	<0.001	-3.3509	-1.0818
1.TV1 - 4.TV2	-3.325	0.43	<0.001	-5.0569	-2.1506
1.TV1 - 1.TV3	-0.413	0.43	0.340	-1.3626	0.4549
1.TV1 - 2.TV3	-0.237	0.43	0.582	-1.1624	0.6401
1.TV1 - 3.TV3	-0.456	0.43	0.292	-1.4134	0.4092
1.TV1 - 4.TV3	-0.219	0.43	0.612	-1.1412	0.6601
2.TV1 - 3.TV1	-0.144	0.43	0.739	-1.0568	0.7408
2.TV1 - 4.TV1	-0.875	0.43	0.046	-1.9127	-0.0179
2.TV1 - 1.TV2	-2.512	0.43	<0.001	-4.0098	-1.5141
2.TV1 - 2.TV2	-2.356	0.43	<0.001	-3.8066	-1.3836
2.TV1 - 3.TV2	-1.838	0.43	<0.001	-3.1316	-0.9316
2.TV1 - 4.TV2	-3.156	0.43	<0.001	-4.8412	-2.0242
2.TV1 - 1.TV3	-0.244	0.43	0.572	-1.1694	0.6334
2.TV1 - 2.TV3	-0.069	0.43	0.873	-0.9733	0.8222
2.TV1 - 3.TV3	-0.287	0.43	0.506	-1.2191	0.5868

Figure 16: Screenshot for the MAM post-hoc for Coloursaturation attribute

Figure 16 shows screenshot of MAM based pairwise product comparisons for attribute `Coloursaturation`. The product differences are shown together with the post-hoc p-value. The new method introduced in Brockhoff et al. (2015) is used for calculating confidence limits. The output is also visualized in barplots.

10.4.5. Post-hoc comparison for each product with the mean of the remaining products

Finally, the table is provided for all attributes, where the MAM based post-hoc comparison of each product with the mean of the remaining products for each attribute is performed. Figure 17

Post-hoc comparison of each product with the mean of the remaining products

	1.TV12.TV13.TV14.TV11.TV22.TV23.TV24.TV21.TV32.TV33.TV34.TV3											
Coloursaturation	-1.22 ***	-1.03 **	-0.88 **	-0.08	1.71 ***	1.54 ***	0.97 **	2.41 ***	-0.77 *	-0.96 **	-0.72 *	-0.98 **
Colourbalance	1.06 .	-0.29	0.35	0.4	-2.36 ***	-0.61	-0.62	-2.59 ***	0.75	1.79 **	1.15 *	0.98 .
Noise	2.31 ***	1.49 **	-2.99 ***	2 ***	-0.12	0.38	-6.52 ***	0.04	1.1 *	1.53 **	-1.44 **	2.22 ***
Depth	-1.4 **	-1.7 ***	-0.34	-0.46	-0.53	0.09	-0.05	-1.09 *	0.25	2.19 ***	1.78 ***	1.27 **
Sharpness	-0.62	-1.75 ***	-0.34	0.26	-1.81 ***	-0.44	-1.1 *	-1.45 **	1.96 ***	3.07 ***	-0.31	2.54 ***
Lightlevel	-2.71 ***	-3.16 ***	-2.84 ***	-2.1 ***	0.8 *	0.4	1.64 ***	1.34 ***	1.3 ***	1.41 ***	0.79 *	3.11 ***
Contrast	-0.91 *	-2.65 ***	-2.19 ***	-0.39	-0.62	0.7 .	-0.16	0.25	1.62 ***	1.81 ***	0.89 *	1.65 ***
Sharpnessofmovement	0.23	-0.9	-1.3 *	1.22 .	-1.05 .	0.64	0.08	0.94	0.52	0.19	-2.62 ***	2.04 **
Flickeringstationary	0.93 .	1.77 ***	-1.93 ***	1.26 *	0.07	0.34	-4.51 ***	-0.35	1.35 **	1.21 *	-0.99 *	0.83 .
Flickeringmovement	1.26 *	0.4	-1.16 .	0.65	0.23	0.4	-1.39 *	-0.28	0.87	0.34	-1.81 **	0.5
Distortion	-0.1	-1.58 **	8.18 ***	0.29	-0.91 .	-1.35 **	-0.63	-0.83	-1.26 *	-0.85 .	0.24	-1.2 *
Dimglasseffect	-0.87 *	-1.18 **	0.8 *	-1.1 **	0.18	-1.16 **	2.2 ***	-0.51	0.4	-0.91 *	3.1 ***	-0.96 **
Cutting	0.96 **	2.89 ***	1.71 ***	0.85 *	0.53	0.66 .	-0.6 .	1.08 **	-1.87 ***	-1.96 ***	-2.12 ***	-2.13 ***
Flossyedges	-1.2 .	1.56 *	-1.57 *	-1.39 *	-1.37 *	1.23 .	-1.57 *	-1.25 .	-0.17	5.55 ***	0.45	-0.28
Elasticeffect	-1.72 ***	-1.61 ***	-0.08	-1.42 ***	-1.27 **	-0.68 .	2.46 ***	-0.89 *	-0.92 *	-0.64	7.71 ***	-0.94 *

Figure 17: Screenshot for the MAM based post-hoc comparison of each product with the mean of the remaining products for each attribute

11. Getting help

To get help on a particular function, e.g. `sensmixed`, you can write `help("sensmixed")` or equivalently `?sensmixed` (or just by typing `sensmixed` in a help tab of RStudio). This works well if you know which function you want help on. And of course you can always use google to search for a particular function.

12. Final remarks

In the analysis of sensory data, quite a lot of random as well as fixed effect can be part of the mixed effects model - this slows down the process of analysis of the data. Depending on the size of the data set, the calculations may take up to few minutes. Whenever one has clicked the *Run analysis* button, in the upper corner a small notification appears, that the calculations

have started and one should wait. If one would like to change the modelling controls of the analysis, then it is important to click the *Run analysis* button in order to run the analysis for the newly selected modelling controls.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. URL: `http://CRAN.R-project.org/package=lme4` r package version 1.0-4.

Brockhoff, P. B., Schlich, P., & Skovgaard, I. (2015). Taking individual scaling differences into account by analyzing profile data with the mixed assessor model. *Food Quality and Preference*, *39*, 156–166.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2015). *shiny: Web Application Framework for R*. URL: `http://CRAN.R-project.org/package=shiny` r package version 0.11.1.

Dahl, D. B. (2014). *xtable: Export tables to LaTeX or HTML*. URL: `http://CRAN.R-project.org/package=xtable` r package version 1.7-4.

Fai, A. H., & Cornelius, P. L. (1996). Approximate f-tests of multiple degree of freedom hypotheses in generalised least squares analyses of unbalanced split-plot experiments. *Journal of statistical computation and simulation*, *54*, 363.

Giesbrecht, F., & Burns, J. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *BIOMETRICS*, *41*, 477–486.

Harvey, W. R. (1975). Least-squares analysis of data with unequal subclass numbers.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2013). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. URL: `http://CRAN.R-project.org/package=lmerTest` r package version 2.0-0.

Kuznetsova, A., Christensen, R. H., Bavay, C., & Brockhoff, P. B. (2015). Automated mixed {ANOVA} modeling of sensory and consumer data. *Food Quality and Preference*, *40, Part A*, 31 – 38. URL: `http://www.sciencedirect.com/science/article/pii/S0950329314001724`. doi:`http://dx.doi.org/10.1016/j.foodqual.2014.08.004`.

Kuznetsova, A., de Sousa Amorim, I., & Brockhoff, P. B. (). Analysing sensory data in a mixed effects model 2 framework using the r package sensmixed. *intended for Food Quality and Preference, Food Qual. Preference*, .

Langsrud, y. (2003). Anova for unbalanced data: Use type ii instead of type iii sums of squares. *Statistics and Computing, Stat. Comput*, *13*, 163–167. doi:`10.1023/a:1023260610025`.

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food*. Springer Science+Business Media, LLC.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010a). *Statistics for sensory and consumer science*. John Wiley and Sons Ltd.

Næs, T., Lengard, V., Blling Johansen, S., & Hersleth, M. (2010b). Alternative methods for combining design variables and consumer preference with information about attitudes and demographics in conjoint analysis. *Food Quality and Preference, Food Qual. Preference*, *21*, 368–378. doi:10.1016/j.foodqual.2009.09.004.

Nofima Mat, N., Ås (2008). Panelcheck software. URL: www.panelcheck.com.

Peltier, C., Brockhoff, P. B., Visalli, M., & Schlich, P. (2014). The mam-cap table: a new tool for monitoring panel performances. *Food Quality and Preference*, *32*, 24–27.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *BIOMETRICS BULLETIN*, *2*, 110–114.

13. Appendix

Some details on how to do test for each of the individual things based on the individually decomposed ANOVA table. All these tests should be used in an "explorative" manner duly taking into account the multiplicity challenges. Also the tests of individuality of disagreement contribution and repeatability level really lies within models that goes beyond the models expressed above where these effects are assumed homogeneous across assessors.

The idea throughout is that each hypothesis test is to be used together with an observed directly interpretable statistic. Then the single overview table of individual effects will include these interpretable statistics - NOT the hypothesis test statistics, but then the result of performing the hypothesis test is indicated below each interpretable statistics.

13.0.6. Assessor individual product discriminability effect

Finally, the individual product effect is tested by the within-individual product difference F-statistic, as also expressed in the section on ANOVA decompositions above:

$$F_i = \frac{SS^{(i)}_{PROD}/(J-1)}{SS^{(i)}_{Error}/(J(K-1))}$$

and a natural statistic to report here is the square root of the individual F_i:

$$\text{Individual product discriminability statistic: } \sqrt{F_i}$$

13.0.7. Assessor scaling

Test for scaling difference:

$$F_i^{Scal1} = \frac{\text{SS}_i^{SCAL}}{\text{SS}^{(i)}_{DIS}/(J-2)} \sim F(1,(J-2))$$

which expressed in the fixed scaling model tests:

$$H_0 : \beta_i = 1 \text{ versus } H_1 : \beta_i \neq 1.$$

This is the same test as one would get by doing a simple regression analysis on the individual data averaged over the replicates $\bar{y}_{ij\cdot}$:

$$F_i^{Scal1} = \frac{(\hat{\beta}_i - 1)^2}{\text{SS}^{(i)}_{DIS}/(J-2)/(SS_{product}/(KI))}$$

The natural statistic to report here is the individual assessor's scaling value:

$$\text{Individual scale statistic: } \hat{\beta}_i$$

These numbers will average to 1.

13.0.8. Assessor disagreement

Test for disagreement: (also using the individual error rather than the pooled error)

$$H_0 : \sigma^2_{DIS,i} = 0 \text{ versus } H_1 : \sigma^2_{DIS,i} > 0.$$

$$F_i^{DIS} = \frac{\mathrm{SS}_i^{DIS}/(J-2)}{\mathrm{SS}_{Error}(i)/(J-2)} \sim F(J-2, J(K-1))$$

The natural statistic to report here is the individual assessor's disagreement standard deviation:

$$\text{Individual disagreement statistic: } \sqrt{\mathrm{SS}_i^{DIS}/(J-2)}$$

This number is comparable with the individual repeatability number given below. The standard deviation here equals the residual standard deviation from a linear regression analysis of the KJ individual scores versus the product average scores. These numbers will average to (approximately) the root of $MS_{Disagreement} \cdot K$.

13.0.9. Assessor repeatability

Test for individual error heterogeneity:

$$H_0 : \sigma_i^2 = \bar{\sigma}^2_{(i)} \text{ versus } H_1 : \sigma_i^2 \neq \bar{\sigma}^2_{(i)}$$

$$F_i^{ERROR} = \frac{\mathrm{SS}_i^{ERROR}}{\frac{1}{I-1}\sum_{\tilde{i}\neq i} \mathrm{SS}_{\tilde{i}}^{ERROR}} \sim F(J(K-1), (I-1)J(K-1))$$

The natural statistic to report here is the individual assessor's error standard deviation:

$$\text{Individual repeatability statistic: } \sqrt{\mathrm{SS}_i^{Error}/(J(K-1))}$$

13.0.10. Assessor level

Test for Assessor effect:

$$F_i^{Ass} = \frac{\mathrm{SS}_i^{ASS}}{\mathrm{MS}_{INT}} \sim F(1, (I-1)(J-1))$$

testing (Expressed in a fixed way)

$$H_0 : \alpha_i = \bar{\alpha}$$

Or maybe slightly more relevant:

$$F_i^{Ass} = \frac{I}{I-1}\frac{\mathrm{SS}_i^{ASS}}{\mathrm{MS}_{INT}} \sim F(1, (I-1)(J-1))$$

testing

$$H_0 : \alpha_i = \bar{\alpha}_{(i)}, \text{ where } \bar{\alpha}_{(i)} = \frac{1}{I-1}\sum_{\tilde{i}\neq i} \alpha_{\tilde{i}}$$

The denominator is the usual interaction mean square:

$$\mathrm{MS}_{INT} = \frac{\mathrm{SS}_{Scaling} + \mathrm{SS}_{Disagreement}}{(I-1)(J-1)}$$

The natural statistic to report here is the individual assessor's average difference to the overall average:

$$\text{Individual level statistic: } \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}$$

These numbers will average to zero.

13.0.11. Assessor correlation

And hence we can similarly test for the sign of the scaling: (negativity and/or positivity)

$$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0$$

by the test:

$$F_i^{Scal0} = \frac{\hat{\beta}_i^2}{\mathrm{SS}_{DIS}^{(i)}/(J-2)/(SS_{product}/(KI))} \sim F(1,(J-2))$$

These tests are using the individual disagreement variability as error term rather than the pooled disagreement across individuals. The latter would be the natural consequence of the classical variance homogeneity assumptions of the linear mixed models, whereas the former more correctly will account for potential heterogeneities between assessors. The natural statistic to report here (in addition to the individual assessor's scaling value) is the correlation between individual scores and average scores:

$$\text{Individual agreement statistic: } r_i = \hat{\mathrm{corr}}\left((\bar{y}_{\cdot 1\cdot},\ldots,\bar{y}_{\cdot J\cdot}),(\bar{y}_{i1\cdot},\ldots,\bar{y}_{iJ\cdot})\right)$$

$$r_i = \hat{\mathrm{corr}}\left((\bar{y}_{\cdot 1\cdot},\ldots,\bar{y}_{\cdot J\cdot}),(\bar{y}_{i1\cdot},\ldots,\bar{y}_{iJ\cdot})\right) = \hat{\beta}_i\sqrt{\frac{SS_{Product}/I}{SS_{Prod}^{(i)}}}$$

These numbers will average to something close to the so-called Cronbach's Alpha.